Estimation automatique des impressions véhiculées par une photographie de visage

Arnaud Lienhard

UNIVERSITÉ GRENOBLE ALPES

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Signal, Image, Parole, Télécoms (SIPT)**

Arrêté ministériel : 7 août 2006

Présentée par

Arnaud LIENHARD

Thèse dirigée par **Alice CAPLIER** et
co-encadrée par **Patricia LADRET**,

préparée au sein du **laboratoire Gipsa-lab**
et de **l'école doctorale d'électronique, électrotechnique, automatique et traitement du signal (EEATS)**

Estimation automatique des impressions véhiculées par une photographie de visage

Thèse soutenue publiquement le **26 novembre 2015**,
devant le jury composé de :

Michèle ROMBAUT
Professeur Université Joseph Fourier, Présidente
Jean-Luc DUGELAY
Professeur EURECOM, Rapporteur
Patrick Le CALLET
Professeur Université de Nantes, Rapporteur
Marina NICOLAS
Ingénieur de Recherche STMicroelectronics, Examinatrice
Alice CAPLIER
Professeur Grenoble INP, Directrice de thèse
Patricia LADRET
Maître de conférence Université Joseph Fourier, Encadrante de thèse

Remerciements

Ces remerciements concluent 3 années passionnantes durant lesquelles j'ai eu le temps d'apprendre énormément de choses, que ce soit sur le plan scientifique ou humain.

Je tiens tout d'abord à remercier mes deux encadrantes, Alice et Patricia, pour m'avoir fait confiance dès le départ et donné la chance de tester tout ce qui me passait par la tête. J'ai particulièrement apprécié votre énergie et votre disponibilité, notamment pour la relecture des différents chapitres de ce manuscrit. Je remercie également tous les membres du jury qui ont accepté d'évaluer ce travail. En plus de l'encadrement scientifique, je tiens à souligner l'aide de Lucia, sans qui je n'aurais probablement jamais réussi à dépasser le stade de l'inscription administrative.

Si ces 3 années ont été formidables, c'est en grande partie grâce à tous mes co-bureaux : Cindy, Céline, Raluca, Pascal, Fahkri, Rafael, Fardin et Tuan. Mais c'est aussi grâce aux parties de cartes et aux différentes soirées avec Quentin, Raphael, la Tim Team, A&R, Taia, Lucas, Guillaume, Miguel, Alexis, Jérémy, Maël, Rémy, Manu, Romain, Edouard, Florian, Benoît, et tous ceux dont je vais me rappeler lorsque ce manuscrit aura déjà été mis en ligne.

Plus particulièrement, un grand merci à :

- Céline et Cindy pour tout ce qui a rendu les journées très courtes : les nombreuses pauses dans le bureau et ailleurs, les problèmes du Gipsa-doc ou la planification des prochaines vacances.
- Quentin et Raphael pour cette pause en Islande en plein milieu de la rédaction. Et pour toutes les soirées incluant des pizzas, une raclette ou le Family's.
- Pascal pour ta bonne humeur contagieuse.
- Raluca pour toutes ces discussions intéressantes sur l'après-thèse.
- Fahkri pour l'invitation à Tunis à l'occasion de ton mariage.
- Tim&Tim pour le lundi soir, les déguisements, le zombicide et les insectes.
- Tous ceux qui coinchent à 80.

Enfin, je tiens à remercier tous ceux qui m'ont aidé pour l'organisation de la soutenance, en particulier mes parents. Merci également à tous mes camarades de l'Ensimag qui ont fait le déplacement pour l'occasion (surtout depuis Paris !). Et bien sûr, merci Hélène, pour me soutenir chaque jour depuis bientôt 5 ans !

Table des matières

Introduction **1**

1 État de l'art **9**

1.1 Introduction 9

1.2 Qualité esthétique d'une photo de visage 11

1.3 Impressions induites par une photo de visage 26

1.4 Résumé des caractéristiques et des algorithmes d'apprentissage 31

1.5 Conclusion 40

2 Analyse des données et apprentissage supervisé **43**

2.1 Introduction 43

2.2 Prétraitements - Gestion des données 46

2.3 Sélection des données pertinentes - L'algorithme Relief 49

2.4 Apprentissage automatique et prédiction 57

2.5 Post-traitement - Fusion des scores 71

2.6 Conclusion 75

3 Estimation de la qualité esthétique d'une photo de visage **77**

3.1 Introduction 78

3.2 Bases d'images annotées 79

3.3 Définition des descripteurs considérés 87

3.4 Définition des régions du visage considérées 97

3.5 Étude de la pertinence des descripteurs et des régions 104

3.6 Étude de la pertinence des algorithmes d'apprentissage supervisé 109

3.7 Estimation de la qualité esthétique et validation du modèle 111

3.8 Comparaison avec l'état de l'art . . . 115
3.9 Application à la recherche et la sélection de photographies de visage . . . 121
3.10 Conclusion . . . 125

4 Estimation de l'impression véhiculée par une photo de visage 129

4.1 Introduction . . . 129
4.2 Bases d'images annotées . . . 131
4.3 Extraction des caractéristiques . . . 135
4.4 Étude de l'influence des caractéristiques . . . 141
4.5 Estimation des impressions de compétence et de sympathie . . . 145
4.6 Comparaison avec l'état de l'art . . . 147
4.7 Fusion des modèles de qualité esthétique et de sympathie . . . 149
4.8 Conclusion . . . 150

Conclusion 155

Bibliographie 166

Table des figures

1 Une même personne peut avoir différents visages 1

2 Classification et régression 4

3 Contraintes sur la position et la taille du visage 5

4 Processus d'apprentissage supervisé 6

1.1 Bilan des caractéristiques et des différentes catégories d'images évoquées 42

2.1 Différentes étapes permettant l'évaluation des images 45

2.2 Étapes liées à l'analyse des images 46

2.3 Illustration de l'algorithme Relief 51

2.4 Machines à vecteur de support 59

2.5 Réseau de neurones artificiel 62

2.6 Prédiction à l'aide d'un réseau de neurones. 62

2.7 Arbres de décision 65

2.8 Construction d'une courbe ROC 69

2.9 Nuages de points non normalisés obtenus par chaque algorithme 72

2.10 Nuages de points normalisés obtenus par chaque algorithme 73

2.11 Nuage de points obtenu par fusion des 4 algorithmes 74

3.1 Exemples d'images présentes dans la base HFS. 82

3.2 Répartition des scores pour la base HFS. 82

3.3 Exemples d'images présentes dans la base PNF. 83

3.4 Répartition des scores pour la base PNF. 83

3.5 Exemples d'images présentes dans la base HFS. 84

3.6 Répartition des scores pour la base FAVA. 84

3.7 Exemples d'images présentes dans la base CUHKPQ 85

3.8 Exemples d'images présentes dans la base Flickr. 86
3.9 Répartition des scores pour la base Flickr. 86
3.10 Exemples de photos évaluées selon le niveau de netteté 90
3.11 Exemples de photos évaluées selon le niveau de textures 91
3.12 Exemples de photos évaluées selon le niveau d'illumination 92
3.13 Exemples de photos évaluées selon le niveau de contraste 93
3.14 Exemples de photos aux couleurs très variées 94
3.15 Exemples de photos évaluées selon les valeurs du canal sombre 96
3.16 Fonctionnement et utilisation des images intégrales 99
3.17 Fonctionnement simplifié de l'algorithme de Viola-Jones 100
3.18 Exemples de détection des attributs faciaux 100
3.19 Bilan des régions considérées dans ce travail 103
3.20 Performances des meilleures caractéristiques en classification 106
3.21 Performances des meilleures caractéristiques en régression 106
3.22 Performances des meilleures couples (C,R) en classification 108
3.23 Performances des meilleures couples (C,R) en régression 108
3.24 Influence des zones de calcul des caractéristiques. 112
3.25 Performances comparées de chaque catégorie de caractéristique. 112
3.26 Nuages de points obtenus par régression sur les bases HFS et FAVA 116
3.27 Valeurs du CCE obtenues sur la base Flickr 118
3.28 Valeurs du MCE obtenues sur la base Flickr 118
3.29 Photographies de bonne qualité esthétique sélectionnées par l'algorithme 123
3.30 Photographies de mauvaise qualité esthétique sélectionnées par l'algorithme . . 124

4.1 Exemples de visages synthétiques 132
4.2 Exemples de photos de la base Karolinska 133
4.3 Exemples de photos de la base HFS_{CS} 134

4.4 Répartition des scores de compétence pour la base HFS_{CS}. 134

4.5 Répartition des scores de sympathie pour la base HFS_{CS}. 134

4.6 Exemple des résultats obtenus par filtrage pour les 6 orientations considérées. . 136

4.7 Positions des points de repère fournis par Betaface et SkyBiometry 137

4.8 Exemples de photos notées selon les impressions de compétence et de sympathie 141

4.9 Classification des visages synthétiques, estimation de la compétence 142

4.10 Classification des visages synthétiques, estimation de la sympathie 142

4.11 Points de repère pertinents selon l'algorithme Relief 143

4.12 Différents nuages de points obtenus pour la base HFS_{CS}. 151

4.13 Application à la sélection de photos d'une personne donnée 152

Liste des tableaux

1.1 Performances des travaux antérieurs sur les bases issues de *Photo.net* 24

1.2 Performances des travaux antérieurs sur les bases issues de *Flickr* 24

1.3 Performances des travaux antérieurs sur les bases issues de *DPChallenge.com* . 25

1.4 Caractéristiques de bas niveau utilisées dans l'état de l'art. 32

1.5 Caractéristiques de niveau intermédiaire utilisées dans l'état de l'art. 34

1.6 Caractéristiques de haut niveau utilisées dans l'état de l'art. 36

1.7 Méthodes d'apprentissage utilisées dans l'état de l'art. 38

3.1 Résumé des différentes bases de photos contenant des visages. 87

3.2 Nombre de détections erronées sur la base LFW 102

3.3 Pertinence des caractéristiques pour les bases d'images HFS et FAVA 105

3.4 Performances des régions en classification et en régression 107

3.5 Couples (C,R) les plus discriminants pour HFS et FAVA 108

3.6 Résumé du tableau 3.5 109

3.7 Performances des différents algorithmes sur les bases HFS et FAVA 110

3.8 Performances de classification binaire sur les 3 bases HFS, FAVA et PNF. . . . 112

3.9 Performances de classification binaire sur les 3 bases HFS, FAVA et PNF, lorsque les photos de score intermédiaire ne sont pas prises en compte. 113

3.10 Performances de classification à 3 catégories sur les bases HFS, FAVA et PNF . 114

3.11 Performances de régression sur les bases HFS, FAVA et PNF 114

3.12 Bilan des performances de classification 119

3.13 Bilan des performances de régression 121

3.14 Temps de calcul moyen par image 123

4.1 Attributs utilisés pour l'estimation des impressions de compétence et de sympathie 138

4.2 Liste simplifiée des attributs utilisés pour l'estimation des impressions. 140

4.3 Attributs les plus discriminants selon l'algorithme Relief 144

4.4 Performances de classification des attributs fournis par chacun des 3 outils . . . 145

4.5 Performances de classification sur la base HFS_{CS} 146

4.6 Performances de régression sur la base HFS_{CS} 146

4.7 Performances de classification sur 200 visages synthétiques 147

4.8 Performances de classification sur 44 photographies 148

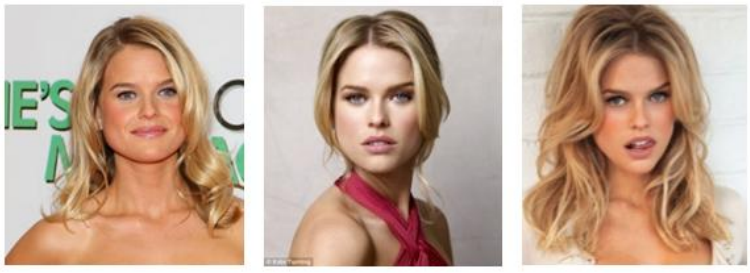
Introduction

Contexte

Avec le développement des moyens de capture et de partage de photographies, nous sommes de plus en plus amenés à diffuser ou consulter des photos. Parmi ces photos, certaines mettent en scène un visage. Ces portraits peuvent par la suite se retrouver sur un réseau social, être utilisés sur des sites de rencontre ou illustrer un CV, ou simplement être triés dans un album de photographies.

Les études en psychologie montrent que l'on se fait très rapidement une opinion sur un individu à partir d'une photo. Les travaux menés par l'équipe d'Alexander Todorov montrent que 100ms peuvent suffire à prononcer un jugement dans le cas de photos de visage [Willis et Todorov 2006]. Ce temps très bref suffit à des observateurs pour évaluer les portraits selon différents critères avec un degré de consensus élevé : une personne souriante est jugée sympathique, un homme en costume est considéré comme quelqu'un de sérieux. Ainsi, même si l'observateur ne s'attarde pas sur la photo qu'il rencontre, il aura tout de même eu le temps de se forger une opinion. Une photo non adaptée à un usage donné peut ainsi rapidement devenir un handicap. Il convient donc de se poser la question : quelles sont les caractéristiques d'une photo de visage réussie, dans le cadre d'une application donnée ? Par exemple, il ne semble pas absurde de penser qu'une photo dégageant une impression de sympathie ou d'amabilité est un atout lorsque l'on souhaite la mettre en avant sur un réseau social. De même, un air sérieux est préférable lorsque la photo doit figurer dans un réseau professionnel. Pour une personne donnée, différentes photos peuvent donc être appropriées à des situations très différentes, comme le montrent les photos présentes sur la figure 1.

FIGURE 1 – Laquelle des 3 photos suivantes est la plus adaptée pour figurer sur un CV ? Un book de mannequin ? Un site de rencontre ?



Dans ce travail, nous nous intéressons dans un premier temps à la qualité esthétique générale d'une photo de visage. **Nous ne cherchons pas à évaluer la beauté du visage**. Nous considérons qu'une photographie est de bonne qualité esthétique lorsque les nombreux aspects qui la caractérisent (cadrage, luminosité, résolution, contraste, flou, équilibre entre les éléments de la composition, etc.) sont de bonne qualité.

La qualité esthétique est un premier critère permettant de trier les photos dans le but de distinguer les photos réussies de celles qui le sont moins. Dans le cadre de la recherche d'images par leur contenu, trier les photos par ordre de qualité esthétique permet à l'utilisateur d'obtenir en premier lieu des photos plus susceptibles de l'intéresser. De plus, il peut être intéressant pour un photographe amateur d'avoir rapidement une idée de la qualité esthétique d'un grand nombre de photos. La présélection rapide des images les plus réussies ou la suppression des images les moins esthétiques pourraient alors être une aide précieuse à la sélection d'une photo adaptée à un usage donné.

Le second point abordé dans ce travail est l'évaluation des différentes impressions que peuvent susciter une photo de visage : cette personne semble-t-elle sympathique ? Menaçante ? Sérieuse ? De la même manière que trier les photos par ordre de qualité esthétique permet de mettre en évidence les photos les plus susceptibles d'intéresser l'utilisateur, ajouter une information sur les impressions véhiculées par les photographies de visage permet de les classer par type d'utilisation : création d'un album de photos, partage entre amis ou en famille, utilisation sur un réseau professionnel.

Sélectionner automatiquement les photos les plus pertinentes est une tâche difficile. De nombreux critères subjectifs entrent dans le processus de décision, et les avis de différentes personnes ne concordent pas toujours. Il est donc nécessaire d'établir un consensus en tenant compte de l'avis de plusieurs observateurs. L'objectif pour l'algorithme est de s'approcher au maximum de ce consensus, même si les prédictions ne s'accordent pas forcément au jugement d'un individu en particulier.

Si l'estimation automatique de la qualité esthétique de photographies quelconques (paysages, portraits, architecture, animaux, etc.) a déjà été abordée sous différents angles, à notre connaissance, peu de travaux concernent la qualité esthétique des photos de visage. La plupart des études portent soit sur l'attractivité du visage, sans tenir compte de la qualité de la photographie, soit sur l'esthétisme de la photographie, sans tenir compte du visage. En partant du principe qu'un paysage n'est pas évalué de la même manière qu'un portrait, les caractéristiques définissant un beau paysage ne sont pas exactement les mêmes que celles définissant une belle photo de visage. L'étude particulière de la qualité esthétique de photographies de visage semble donc pertinente : celle-ci n'a pas été réalisée auparavant et les travaux réalisés sur la qualité esthétique de photographies en général ne s'adaptent pas forcément au cas des photos de visage.

De plus, très peu de travaux ont été menés afin d'estimer automatiquement les différentes impressions (sympathie, menace, compétence, etc.) véhiculées par une photographie. Des méthodes existent, mais les expériences sont essentiellement réalisées sur des bases de visages synthétiques (visages générés par ordinateur) dont les traits sont exagérés. Récemment, Mazza et al. ont réalisé des expériences afin de déterminer les caractéristiques (âge, sexe, vêtements, etc.) influant sur les impressions dégagées par les photographies. Toutefois, les photos ont été annotées par des humains alors que nous cherchons à extraire toutes les informations automatiquement. L'étude des impressions dégagées par une photographie de visage est donc également pertinente et la création d'algorithmes permettant de les évaluer automatiquement est un objectif qui n'a pas encore été atteint.

Objectifs de ce travail

Dans ce travail, nous cherchons à déterminer les caractéristiques déterminant la réussite d'une photo de visage dans le cadre d'une application précise. Pour cela nous étudions dans un premier temps la qualité esthétique d'une photo de visage, puis nous cherchons à estimer les impressions de compétence et de sympathie véhiculées par une photo de visage. Nous avons choisi de nous limiter à ces deux impressions car celles-ci nous semblent particulièrement adaptées aux utilisations les plus courantes d'une photo de visage. En effet, sélectionner automatiquement des photographies présentant un visage sympathique permet de choisir rapidement des photographies appropriées à un partage en famille ou entre amis (réseau social, album de photos). A l'inverse, conserver les photographies où le visage semble compétent (sérieux, digne de confiance) permet de choisir des photos adaptées à un usage professionnel (CV, carte de visite, réseau social professionnel).

Classification de photographies

L'objectif étant de sélectionner les photos les plus réussies adaptées à une certaine application, nous cherchons dans un premier temps à distinguer une photographie réussie d'une photographie ratée. Ce problème est un problème de classification binaire : les photos sont soit considérées comme réussies (convenant à une application donnée), soit comme étant ratées (ne convenant pas à l'application). La classification correspond ainsi au calcul d'un modèle permettant de prédire une variable discrète (une catégorie) à l'aide d'un ensemble d'informations extraites sur la photo. Du point de vue de l'utilisateur, la classification permet donc de trier rapidement les photos.

Dans la plupart des bases de données utilisées pour la classification, les images sont initialement associées à des scores de vérité terrain correspondant à la moyenne des scores fournis par des observateurs humains. Pour adapter ces informations à un problème de classification, nous devons définir deux catégories d'images. Ainsi, la moitié des photos dont le score est inférieur au score médian est associée à une catégorie, et l'autre moitié à l'autre catégorie. Une des limites imposées par le choix de deux catégories est le fait qu'il est très difficile d'évaluer les photos dont le score est proche du score médian. Dans le cas d'une situation réelle, il existe en effet tout un ensemble d'images intermédiaires, qu'il est difficile de classer dans l'une ou l'autre catégorie (photographies "moyennes").

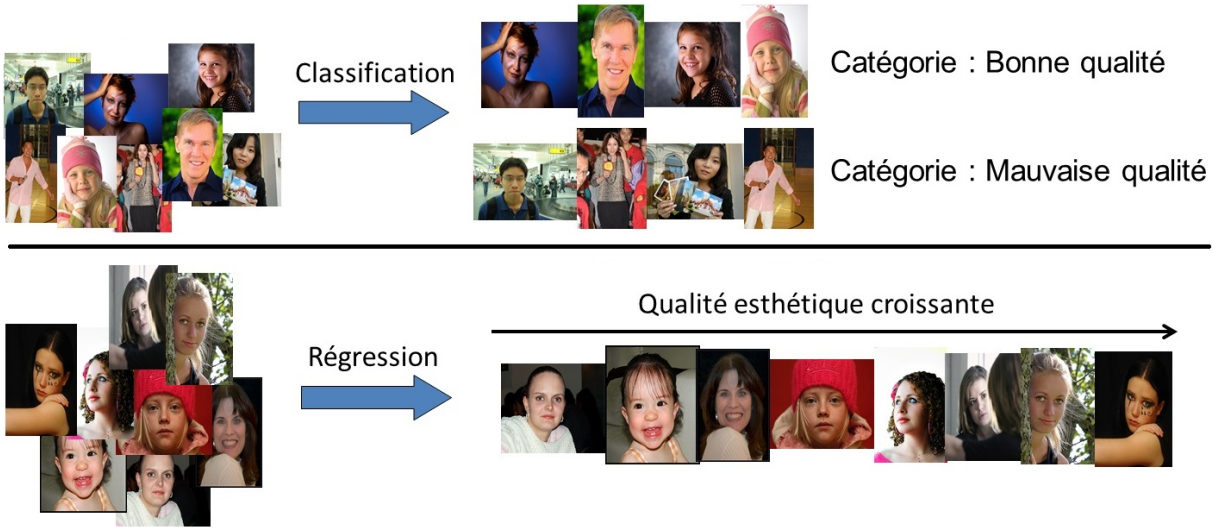
Pour tenir compte de ces photos moyennes, il est possible d'ajouter une classe intermédiaire associée à ces photos dans le modèle. La répartition peut alors se faire en trois tiers égaux (même nombre d'images dans chaque catégorie) ou selon des valeurs précises (scores entre 0 et 3, entre 3 et 7, entre 7 et 10 par exemple). Ainsi, des photos réussies (classe 1) peuvent être confondues avec des photos moyennes (classe 2), et des photos moyennes avec des photos ratées (classe 3), mais il est plus rare que les photos de la classe 1 soient confondues avec celles de la classe 3, sous réserve que le modèle de classification soit pertinent.

Attribution d'un score à une photographie

La régression correspond à la réalisation d'un modèle permettant de prédire une variable continue (ici un score moyen) à l'aide d'un ensemble d'informations extraites sur la photo. Nous cherchons ici à attribuer automatiquement à chaque photo une note (de qualité esthétique, d'impression de sympathie, etc.). Ce problème revient à faire de la classification dont le nombre de classes est infini. Pour la régression, il n'est pas nécessaire de transformer les scores fournis par les utilisateurs, l'objectif étant d'utiliser directement ces scores et de les définir comme étant l'objectif à atteindre par le modèle. Celui-ci est ainsi d'autant plus performant que les scores prédits s'approchent des scores fournis par les observateurs humains.

Créer un modèle de régression permet aux utilisateurs du modèle d'obtenir des informations sur une image afin de prendre une décision particulière. Par exemple, récupérer les 10 photos dont le score est le plus élevé laisse à l'utilisateur la possibilité de choisir parmi cette présélection celle qui lui convient le mieux. La régression permet également de trier les photographies par ordre de pertinence. La figure 2 résume les différences entre classification et régression.

FIGURE 2 – La classification consiste à classer les images dans la catégorie qui lui correspond : images de bonne ou de mauvaise qualité esthétique. La régression consiste à attribuer un score aux images, de façon à pouvoir les ordonner.



Contraintes sur les images

Jusqu'à présent, nous avons discuté de photos de visage sans poser de définition particulière. Cette section pose ainsi les contraintes que nous fixons aux photos dans notre étude. Une première contrainte évidente sur les photos que nous considérons est la présence d'un visage. Différentes catégories de photos peuvent intégrer des visages, celles-ci peuvent être des photos de groupe (plusieurs visages présents), des portraits représentant le corps d'une personne, ou

encore des photos dont le visage est la composante essentielle de la photo.

Nous nous intéressons spécialement à ce dernier type de photographie, et dans ce document, nous désignons par photo de visage une photo dont le sujet est un visage centré dans l'image, et tronqué au niveau des épaules et du haut de la tête. Nous supposons également que le visage est entièrement visible, c'est-à-dire qu'il est faiblement orienté et qu'il n'y a pas d'occultation (les yeux, le nez et la bouche sont visibles). Ces dernières limites sont essentielles pour les modèles que nous proposons, car nous utilisons des algorithmes de détection de visages présentés de face, et que nous extrayons des informations relatives aux yeux et à la bouche.

En pratique, nous considérons que ces conditions sont satisfaites lorsque :

1. Un unique visage est détecté dans l'image,
2. La taille et la position du visage dans la photo respectent les conditions définies sur la figure 3.

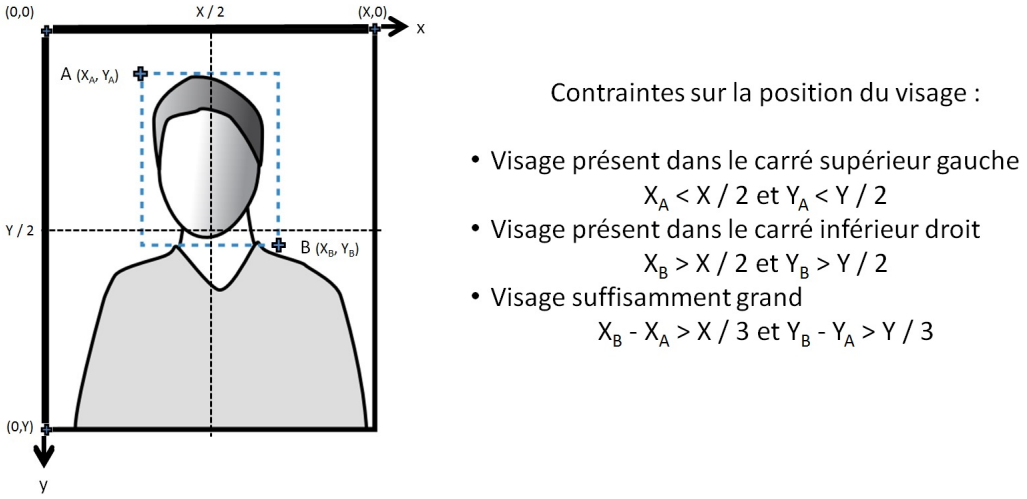


FIGURE 3 – Contraintes sur la position et la taille du visage pour que la photo soit considérée comme une photo de visage dans nos travaux.

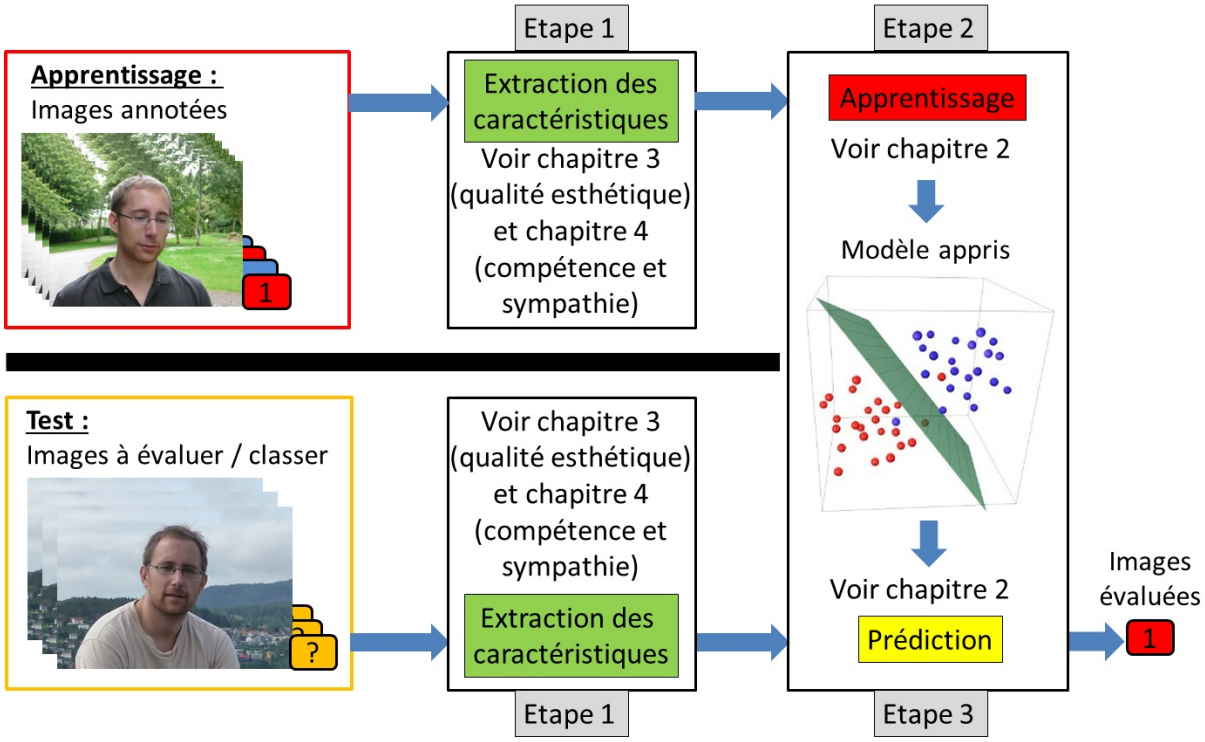
En fixant ces contraintes, particulièrement strictes, nous espérons éviter un certain nombre de biais introduits dans les évaluations de photos contenant des visages. Par exemple, lorsque plusieurs visages sont présents, les relations entre chaque visage (position, expressions) influent fortement sur notre perception de la photo ([Li et al. 2010a ; Xue et al. 2013]). De même, lorsque le visage n'est pas entièrement visible ou suffisamment grand, celui-ci n'est plus forcément le sujet principal de l'image et est par conséquent plus difficile à analyser.

Apprentissage supervisé

Que ce soit pour l'estimation de la qualité esthétique, de la compétence ou de la sympathie, nous cherchons d'abord à déterminer les caractéristiques discriminantes d'une photo de visage, puis à prédire automatiquement la qualité esthétique, la compétence et la sympathie

à partir de ces caractéristiques. Afin de réaliser cet objectif, l'approche que nous considérons comprend 3 étapes, qui définissent un processus d'apprentissage supervisé. Ces étapes sont résumées sur la figure 4. La première consiste en l'extraction de caractéristiques discriminantes (par exemple : "L'image est-elle floue ? Le visage est-il souriant ?"), que nous calculons sur un ensemble conséquent de photographies annotées au préalable par des humains selon le critère étudié (par exemple : "Cette photo est-elle de bonne ou de mauvaise qualité esthétique ? La personne représentée semble-t-elle sympathique ?"). Nous utilisons ensuite un algorithme d'apprentissage supervisé afin d'associer les valeurs des caractéristiques calculées aux annotations des photographies. Une fois cet apprentissage terminé, il est possible de prédire si une nouvelle photographie est réussie ou non, selon le critère considéré.

FIGURE 4 – Schéma simplifié du processus d'apprentissage supervisé. Des caractéristiques sont extraites sur les photos de la base d'apprentissage (étape 1), ce qui permet de construire un modèle de classification ou de régression (étape 2). Ce modèle permet de faire de la prédiction à partir des caractéristiques extraites sur les images de test (étape 3).



L'objectif pour nos algorithmes est de reproduire les évaluations et les jugements humains. Ainsi, si la plupart des humains considèrent qu'une photo de visage est de bonne qualité esthétique, l'algorithme est considéré comme performant s'il est capable d'évaluer la photo comme étant de bonne qualité esthétique. De même, si un homme en costume semble très compétent pour les humains, l'algorithme doit également associer la présence d'un homme en costume à un score de compétence élevé. Rappelons que **nous ne cherchons pas à évaluer la compétence ou la sympathie réelle d'une personne**, mais uniquement l'impression véhiculée par sa photo. La figure 1 montre en effet que plusieurs photographies d'une même personne peuvent suggérer des impressions totalement différentes, notre objectif étant de classer auto-

matiquement ces photos selon les impressions de compétence ou de sympathie dégagées par la personne.

Pour résumer, dans ce travail, nous cherchons à :

1. Étudier quelles sont les caractéristiques discriminantes de la qualité esthétique de photographies de visage afin de créer un algorithme d'estimation automatique de la qualité esthétique de ce type de photographies.
2. Étudier quelles sont les caractéristiques discriminantes pour les impressions de sympathie et de compétence véhiculées par une photo de visage afin de créer un algorithme d'estimation automatique de ces impressions.
3. Proposer une méthode de fusion de ces estimations, permettant de trier automatiquement les photographies dans le cadre d'une application précise. Par exemple extraire une photo de bonne qualité esthétique et dégageant une impression de compétence pour la mettre sur un CV.

Plan et contributions de ce document

Le chapitre 1 référence les travaux précédents sur le sujet étudié, et liste notamment les différentes méthodes employées pour l'estimation de la qualité esthétique de photographies. Il renseigne sur les différentes caractéristiques considérées jusqu'à présent. Ce premier chapitre résume également les études réalisées pour l'estimation des impressions de compétence et de sympathie. Nous montrons que ces problématiques (estimation de la qualité esthétique, compétence et sympathie) sont récentes et coïncident avec l'apparition et le développement d'Internet et des appareils photo numériques.

Le chapitre 2 présente les différents outils que nous employons afin d'évaluer la qualité esthétique et les impressions de compétence et de sympathie. En effet, nous avons défini un cadre de travail générique permettant d'obtenir des résultats pertinents pour ces deux problématiques. Nous détaillons dans ce chapitre les différentes méthodes employées pour l'analyse des caractéristiques discriminantes ainsi que les algorithmes d'apprentissage supervisé permettant la création des modèles d'estimation. Dans ce chapitre, nous améliorons les méthodes actuelles par les contributions suivantes :

1. L'adaptation de l'algorithme de sélection de caractéristiques Relief à nos problématiques, ce qui augmente la précision de nos prédictions tout en réduisant les temps de calcul des caractéristiques et d'apprentissage.
2. L'utilisation de 4 algorithmes d'apprentissage différents (SVM, ANN, RF, GBT), qui permet de s'assurer de la robustesse des prédictions.
3. La fusion des sorties proposées par les 4 algorithmes d'apprentissage, qui améliore également la robustesse et de la précision des prédictions.

Enfin, **les chapitres 3 et 4** traitent respectivement des estimations de la qualité esthétique et des impressions de compétence/sympathie. Dans ces deux chapitres, nous définissons un jeu de caractéristiques approprié à chacune des problématiques, puis nous utilisons les algorithmes

définis dans le chapitre 2 afin de mettre en évidence les caractéristiques les plus pertinentes et de proposer des modèles d'estimation précis.

Les contributions de ce travail à l'estimation de la qualité esthétique de photographies de visage (**chapitre 3**) sont les suivantes :

1. La création de deux bases de photos de visage distinctes, pour lesquelles les images sont évaluées selon leur qualité esthétique. La première base (250 images) contient des images évaluées dans un environnement très contrôlé, tandis que la seconde comprend plus de 28000 photos de visage récupérées automatiquement sur le site [*PhotoNet*].
2. La définition de différentes caractéristiques décrivant des informations globales (netteté, illumination, contraste, couleurs) sur une région particulière. Ces caractéristiques sont choisies pour leur capacité à discriminer des images de qualité esthétique différentes.
3. Le calcul de ces caractéristiques sur différentes régions d'une photo de visage, comprenant les régions définies par les yeux ou la bouche. La prise en compte de ces informations locales améliore significativement les performances des estimations de l'état de l'art.
4. L'analyse des caractéristiques et des régions du visage pertinentes. Nous montrons entre autre que l'extraction de caractéristiques uniquement sur la région des yeux, très riche en information, permet d'obtenir de très bonnes performances alors que la région est très petite par rapport à la taille de l'image. Ce résultat permet d'accélérer significativement l'évaluation dans le cadre d'une utilisation en temps réel.
5. L'application de la méthode proposée à la sélection automatique des photographies dont la qualité esthétique est la plus élevée.

Enfin, les contributions de ce travail à l'estimation des impressions de sympathie et de compétence (**chapitre 4**) sont les suivantes :

1. La création d'une base de 140 photos de visage, annotées selon les impressions compétence et la sympathie dégagées par les photos.
2. La définition et l'utilisation d'attributs de haut niveau d'interprétation (par exemple : présence de sourire, ouverture des yeux, sexe de la personne) afin de créer des modèles d'évaluation des impressions suggérées par une photo de visage. Nous montrons qu'avec ces attributs, une amélioration significative des performances de l'état de l'art peut être obtenue.
3. Une démonstration de la faisabilité du tri automatique de photographies en fonction de l'impression de sympathie véhiculée par une photo de visage, pour une personne donnée.
4. L'application de la méthode proposée à la sélection automatique des photographies dont la qualité esthétique ainsi que l'impression de sympathie véhiculées par la photographie sont élevées.

Chapitre 1

État de l'art

Sommaire

1.1 Introduction 9
1.2 Qualité esthétique d'une photo de visage 11
1.2.1 Lien entre qualité et qualité esthétique d'une photographie 11
1.2.2 Étude de la qualité esthétique de photographies 14
1.2.3 Étude de la qualité esthétique de photos de visage 18
1.2.4 Principales bases de photos évaluées 22
1.3 Impressions induites par une photo de visage 26
1.3.1 Importance des premières impressions et du contexte 26
1.3.2 Méthodes d'estimation des traits de caractère 27
1.3.3 Principales bases de visages annotées 29
1.4 Résumé des caractéristiques et des algorithmes d'apprentissage . . . 31
1.4.1 Caractéristiques considérées 31
1.4.2 Algorithmes d'apprentissage automatique 37
1.5 Conclusion 40

1.1 Introduction

Les premiers travaux ayant pour thème la qualité esthétique de photos remontent au début des années 2000. Ces années coïncident avec l'apparition d'Internet ainsi que d'appareils photo numériques dans un grand nombre de foyers. Il devient alors intéressant de pouvoir filtrer automatiquement les images que l'on souhaite partager ou non. Ce filtrage peut s'effectuer lors de l'acquisition de la photo, par exemple en adaptant les paramètres physiques de l'appareil photo : déclenchement automatique du flash, mise au point automatique afin d'éviter la prise de photos floues, détection de sourire, etc. Il est également possible de trier les photos à conserver après l'acquisition de ces dernières, par exemple en choisissant celles présentant les plus beaux sourires. Dans ce travail, nous nous plaçons dans ce dernier cas de figure, l'objectif étant d'évaluer et de classer automatiquement les photos de manière à éviter de consulter manuellement chacune des images.

L'appréciation que chacun peut se faire d'une photographie est évidemment subjective. Toutefois, il existe un certain nombre de critères pour lesquels un consensus peut être trouvé :

image floue (flou involontaire ou artistique), sur ou sous-exposée, mauvais cadrage, etc. Ces critères définissent des photos de mauvaise qualité esthétique et peuvent être détectés automatiquement à l'aide des méthodes qui sont présentées dans ce chapitre. Il est important de noter que les caractéristiques permettant d'apprécier ou non une photo dépendent du type d'image considéré : un beau paysage ne suit pas les mêmes règles qu'un beau portrait. Les critères définissant une photo réussie peuvent également varier selon des objectifs particuliers. La photo est-elle destinée à un partage avec des amis ? À un usage professionnel ?

Dans la section 1.2, nous abordons les méthodes permettant l'estimation de la qualité esthétique de portraits. Nous étudions tout d'abord un premier critère relatif à l'esthétisme d'une image : sa qualité. Celle-ci correspond à sa résolution, sa netteté, son rendu chromatique, etc. Ces aspects de l'image ont une influence directe sur sa qualité esthétique. Nous considérons ensuite différents travaux directement liés à l'évaluation de l'esthétisme d'images indépendamment de leur type (portrait, paysage, architecture, etc.), puis les techniques prenant en compte les critères propres aux photos de visage. Enfin, nous décrivons brièvement les bases de données utilisées dans la littérature pour tester les différentes méthodes.

L'appréciation d'une image s'effectue dans un contexte particulier. Ce contexte varie selon l'utilisation de la photographie : un portrait destiné à la création d'un album partagé en famille ne sera pas forcément considéré pour un usage professionnel, et inversement. Dans certains cas, l'image considérée doit dégager une impression de sympathie (partage sur un réseau social), dans d'autres c'est un air sérieux et compétent qui est préféré (usage professionnel). Dans la section 1.3, nous étudions les différents travaux permettant l'estimation automatique de différents traits subjectifs tels que l'impression de sympathie ou de compétence dégagée par une photo de visage. Ces travaux reposent essentiellement sur l'étude de points d'intérêt dans le visage : contours des yeux, de la bouche, position des sourcils. Ces points traduisent la forme générale du visage ainsi que des expressions faciales (joie, colère, etc.), fortement corrélées aux évaluations de sympathie, compétence, etc. La plupart des méthodes existantes sont validées à l'aide de bases de visages synthétiques, décrites dans la dernière partie de la section.

Un récapitulatif des métriques utilisées dans l'état de l'art est présenté dans la dernière section de ce chapitre. Nous distinguons les mesures extraites directement à partir des pixels de l'image (mesures de bas niveau) et les mesures nécessitant un niveau d'abstraction supplémentaire par rapport aux pixels de l'image (mesures de haut niveau). Ces dernières correspondent à des attributs proches du jugement humain : la personne est-elle souriante ? Les mesures relatives à la composition de l'image ou à la présence d'un objet particulier (dans notre cas un visage) sont également décrites.

1.2 Qualité esthétique d'une photo de visage

1.2.1 Lien entre qualité et qualité esthétique d'une photographie

Comme nous l'avons évoqué dans l'introduction, dans ce travail nous considérons qu'une photographie est de bonne qualité esthétique lorsque les nombreux aspects qui la caractérisent (cadrage, luminosité, résolution, contraste, flou, équilibre entre les éléments de la composition, etc.) sont de bonne qualité. Certains de ces éléments traduisent également la qualité visuelle d'une photo : image nette, contrastée, non bruitée. La qualité esthétique inclut ainsi non seulement des critères relatifs à la qualité de l'image, mais comprend également les différents effets permettant la mise en valeur du sujet (composition de l'image, orientation de l'éclairage, choix des couleurs) ainsi que la capacité de l'image à capter notre attention. Les notions de qualité photo et de qualité esthétique sont donc étroitement liées. Toutefois, une bonne qualité n'est pas toujours un gage d'esthétisme, et réciproquement.

Dans cette section, nous traitons brièvement des méthodes d'estimation de la qualité d'une photographie, puis des premiers travaux dont l'objectif est de distinguer des photographies professionnelles de photographies amateurs (pas forcément de visage). Cet objectif est très proche du nôtre, qui est l'estimation de la qualité esthétique d'une photographie de visage.

Qualité d'une photographie

L'estimation automatique de la qualité de photographies a fait l'objet de nombreux travaux lorsque les appareils photos numériques se sont imposés dans les foyers. Les différents modèles établis dotent alors les appareils numériques d'algorithmes capables d'adapter la luminosité ou la mise au point (autofocus). Dans le cas de photos de visage, estimer la qualité permet par exemple d'adapter les algorithmes de reconnaissance faciale au niveau de dégradation de l'image [Fiche et al. 2010 ; Liao et al. 2012].

De nombreuses métriques de comparaison entre les photos transformées (compressées, bruitées) et originales existent : $PSNR$, $SSIM$. Celles-ci présentent l'inconvénient de nécessiter une image de référence (non transformée) dont nous ne disposons que très rarement. Il existe des métriques d'estimation directe (sans image de référence) de qualité d'image, comme l'algorithme présenté récemment par [Liu et al. 2014], dont le principe repose sur le calcul des entropies spatiales et spectrales de l'image. Dans ce travail, nous n'utilisons pas d'algorithmes d'estimation de compression ou de bruit, car ceux-ci sont souvent liés à des formats ou des conditions de capture spécifiques. De plus, nous espérons tenir compte de ces artéfacts (bruit, compression) dans d'autres caractéristiques décrites dans ce document (netteté, textures).

Vers la qualité esthétique d'une photographie

[Tong et al. 2004] puis [Ke et al. 2006] présentent tous deux des méthodes permettant la classification de photographies en distinguant des photos prises par des professionnels des

photos amateurs. Cet objectif se situe à mi-chemin entre l'évaluation de la qualité d'une image et l'évaluation de son esthétisme, car un photographe professionnel a tendance à produire des images esthétiques de bonne qualité, tandis que les clichés amateurs présentent souvent des défauts (mauvaise qualité, faible esthétisme). La principale différence entre les deux méthodes provient des métriques considérées. Tong et al. utilisent deux types de mesures. Dans un premier temps, différentes mesures globales (représentées par 8 valeurs réelles) sont calculées sur l'image : estimations de netteté, de contraste, de coloration, ainsi que des statistiques calculées sur la carte de saillance issue de la méthode présentée par [Ma et al. 2002]. Ensuite, différentes caractéristiques basées sur des histogrammes de couleur, des décompositions en ondelettes, des transformées de Fourier ou encore sur des filtres de Canny ou Sobel sont calculées. Un total de 21 types de caractéristiques est calculé, représentant 846 valeurs réelles décrivant chaque image. De leur côté, Ke et al. n'utilisent qu'un nombre très réduit de mesures. Ces mesures quantifient la distribution des contours et des couleurs, le nombre de couleurs, la netteté (distribution des contours et estimation du flou), le contraste et la luminosité. Chaque image est ainsi représentée par seulement 7 valeurs réelles, contre 846 pour Tong et al.

Pour tester la pertinence de ces choix de caractéristiques, la méthode employée consiste à apprendre un modèle à partir d'une base de photos évaluées par des humains à l'aide d'un algorithme d'apprentissage automatique. L'objectif de ces travaux est de distinguer automatiquement les photos de professionnels des photos amateurs. Pour cela, Tong et al. considèrent environ 30000 photos classées dans chaque catégorie en fonction de leur provenance (sites professionnels ou photos d'amateurs), tandis que Ke et al. récupèrent 60000 photos sur le site [*DPChallenge*]. Ces photos sont notées par un nombre conséquent d'internautes selon leur qualité esthétique. Seules les 10% de photos (soit 6000) dont le score est le plus faible ainsi que les 10% dont le score est le plus élevé sont considérées. Le premier groupe est alors défini comme étant des photos amateurs, tandis que le second groupe correspond à des photos dont la qualité correspond à des clichés de professionnels. Les résultats de la classification montrent que les deux méthodes produisent des résultats équivalents pour une même base d'images et la même méthode d'apprentissage, bien que le nombre de caractéristiques utilisées par Ke et al. soit largement inférieur. Cela signifie qu'un très grand nombre de caractéristiques (846 pour Tong et al.) n'est pas nécessaire, et qu'il est possible de se limiter à un faible nombre de caractéristiques discriminantes (7 pour Ke et al.).

Intérêt des différentes régions de l'image

Si les méthodes précédentes présentent quelques mesures permettant d'appréhender la notion d'avant / arrière-plan à l'aide de l'étude de la distribution des contours et des couleurs, la plupart des caractéristiques calculées ne concernent que l'image dans son ensemble. Nous allons maintenant voir que la localisation du sujet dans une photo afin de distinguer les différentes régions de l'image permet d'améliorer l'évaluation de la qualité d'image en étudiant séparément chaque région. Une technique permettant de détecter le sujet de la photographie est décrite par [Luo et Tang 2008]. Encore une fois, l'objectif est de distinguer les photos de professionnels des photos d'amateurs. Le principe de la méthode présentée par [Luo et Tang 2008] est de définir la région contenant le sujet comme la zone où l'image est la plus

nette. Il est ensuite possible de calculer différentes caractéristiques sur chaque région (sujet, arrière-plan). En plus de caractéristiques similaires à celles présentées par [Tong et al. 2004] et [Ke et al. 2006], il devient possible d'étudier la composition de l'image : [Luo et Tang 2008] considèrent également la position du sujet dans l'image. Afin d'évaluer les performances de cette nouvelle méthode, le même protocole expérimental que celui présenté par [Tong et al. 2004] et [Ke et al. 2006] est employé : apprentissage supervisé à partir d'une base de données puis classification. Les expériences montrent que pour un algorithme et une base de données identiques, les résultats obtenus en distinguant le sujet de l'arrière-plan [Luo et Tang 2008] dépassent sensiblement les résultats obtenus en étudiant uniquement l'image dans sa globalité [Ke et al. 2006]. Cette idée sera mise en avant dans ce document, dans le cas particulier des photos de visage : pour en estimer correctement la qualité, il est au minimum nécessaire de localiser le visage dans la photo.

Il existe d'autres méthodes permettant de tenir compte des différentes régions d'une image. Par exemple, l'utilisation d'une carte de saillance permet de déterminer les régions qui attirent le regard. Les cartes de saillance mettent en avant les zones dont les couleurs sont les plus vives, les contours fortement marqués, et le contraste élevé. Toutefois, dans le cas d'un portrait, ou plus généralement dans le cas de photos contenant des personnes, ce sont avant tout les visages qui attirent l'attention, au point que la détection de visages est généralement intégrée lors de la construction de ces cartes [Marat et al. 2013]. Étant donné que nous ne nous intéressons qu'aux photos de visage, nous n'intégrons pas de cartes de saillance dans nos modèles et nous nous limitons à la détection du visage et de ses contours.

Conclusion

Les auteurs des travaux présentés ici ne distinguent pas la qualité d'une image et sa qualité esthétique : l'objectif est de distinguer des photographies de professionnels de photographies d'amateurs, et non d'estimer la qualité ou la qualité esthétique d'une image. Toutefois, les caractéristiques utilisées dans ces travaux présentés encodent essentiellement des informations concernant la qualité d'une image (netteté, contraste), et finalement peu d'informations sur la qualité esthétique (composition de l'image via la distinction de l'avant et de l'arrière-plan). Afin d'estimer la qualité esthétique, il est nécessaire de considérer également d'autres caractéristiques : son sujet, sa composition, ses couleurs, etc.

Tous les travaux évoqués, ainsi que ceux présentés dans ce document, ont pour objectif d'évaluer les images uniquement grâce aux informations obtenues à partir des pixels. Nous n'utilisons pas les méta-données incluses dans certains fichiers, car un nombre conséquent de photos circulent sans ces informations. Enfin, nous ne considérons pas les conditions de visualisation de l'image. S'il est évident que le confort de visualisation (écran de bonne qualité, distance à l'écran adaptée) joue un rôle dans la perception de l'esthétisme d'une image, l'objectif de ce travail est de donner une estimation de la qualité esthétique intrinsèque de l'image.

1.2.2 Étude de la qualité esthétique de photographies

Les travaux décrits ici sont plus récents et se basent essentiellement sur ceux évoqués en 1.2.1. La méthode générale reste la même : une base de photos est constituée, des caractéristiques sont calculées, et un apprentissage automatique supervisé est réalisé.

Premiers travaux

Les travaux présentés par [Datta et al. 2006] constituent un premier pas conséquent vers une estimation automatique de la qualité esthétique de photographies. Les 3500 photos utilisées pour leurs expériences proviennent du site [*PhotoNet*], un site de partage de photographies où les photos sont associées à deux notes données par les internautes, correspondant à leur intérêt et leur esthétisme. Seules les notes de qualité esthétique sont considérées dans cet article. Une cinquantaine de caractéristiques sont extraites et témoignent à la fois de la qualité photo (éclairage, textures, couleurs) mais également sa composition (convexité des formes, règle des tiers, différences de texture entre le centre et les bords) et d'autres critères tels que la taille de l'image, la familiarité (distance par rapport à un ensemble de photos connues), etc. Seules les images possédant des scores très élevés ou très faibles sont conservées, l'objectif étant de distinguer les photos les plus esthétiques des photos moins réussies. Les résultats de classification oscillent autour de 70% de bonne classification, ce qui est largement au-dessus du hasard (50%). Il a également été tenté de faire une régression afin de prédire automatiquement les scores des photos : les résultats de la prédiction sont également plus performants que le hasard.

Descripteurs génériques

Des informations sur une image peuvent être obtenues à partir de l'extraction de descripteurs qui témoignent de la structure locale de l'image, à l'instar des descripteurs d'image génériques utilisés par [Marchesotti et Perronnin 2011]. L'idée est d'extraire les informations à partir de descripteurs SIFT [Lowe 1999] ou SURF [Bay et al. 2006] sur l'image, puis de constituer un dictionnaire dont chaque mot correspond à une instance possible du descripteur. L'histogramme des fréquences de chaque mot est alors utilisé pour décrire l'image, et un algorithme d'apprentissage peut être appliqué afin de classer les images selon leur histogramme et leur qualité esthétique. Ces descripteurs ont l'avantage de s'adapter à n'importe quel type d'image et d'être utilisés dans d'autres domaines du traitement de l'image. Toutefois, certaines informations sont manquantes. Ainsi, pour tenir compte de la composition des images (règle des tiers, avant/arrière-plan), une solution est de construire des histogrammes dans chaque région ou à différentes échelles de l'image. De plus, la plupart des descripteurs classiques (SIFT, SURF) reposant sur des images en niveau de gris, des mesures de couleur doivent être ajoutées. En tenant compte de cela, et en remplaçant la construction du dictionnaire par une distribution basée sur des vecteurs de Fisher, [Marchesotti et Perronnin 2011] obtiennent des performances significativement supérieures à celles des travaux de [Ke et al. 2006] et [Datta et al. 2006], avec des taux de bonne classification passant d'environ 75% à

pas loin de 90% pour une même base d'images. Cela signifie que l'utilisation de descripteurs génériques est une alternative pertinente permettant de discriminer efficacement des images de qualité esthétique très différente. Nous avons cependant fait le choix de ne pas utiliser ces descripteurs, car les histogrammes décrivant les valeurs calculées sont difficiles à interpréter visuellement.

Mesures de haut niveau

Plutôt que d'extraire des informations locales et de "bas niveau" (information extraite directement par l'étude des pixels), [Dhar et al. 2011] considèrent des mesures de "haut niveau" (informations sémantiques, proches de la compréhension humaine, difficiles à extraire directement des pixels) afin de tenir compte de caractéristiques abstraites telles que le type d'image considéré (visages, animaux, objets), l'environnement entourant le sujet (intérieur / extérieur, météo) et la composition de l'image. Ce type de critères nécessite un apprentissage préalable et donc un travail plus conséquent en amont. Typiquement, détecter la nature et la position de l'objet pris en photo requiert une méthode de détection efficace. [Dhar et al. 2011] montrent dans leurs expériences que l'utilisation conjointe de mesures de bas et de haut niveau améliorent de manière significative les performances de classification par rapport aux travaux de [Ke et al. 2006].

Composition de l'image

Des métriques permettant d'appréhender la composition de photographies sont présentées par [Ng et al. 2009] : détection de visage, étude des lignes estimées par transformée de Hough, différence d'entropie entre la gauche et la droite de l'image, etc. En ajoutant ces caractéristiques relatives à la composition à des outils d'estimation de qualité photo (voir 1.2.1), il est possible d'arriver à un taux de distinction entre photos de bonne et de mauvaise qualité esthétique de l'ordre de 80%, contre 72% pour [Ke et al. 2006] sur la même base de photos. L'étude des performances de chaque caractéristique montre que l'existence d'un point de fuite (déterminé par l'intersection des lignes obtenues par transformée de Hough) ainsi que l'aire de la région d'intérêt (obtenue par l'utilisation de cartes de saillance) sont les mesures liées à la composition les plus pertinentes. Une application possible de l'étude de la composition de l'image est le recadrage automatique, en découpant par exemple les régions qui ne présentent que peu d'intérêt.

Estimation du sujet de la photo

L'estimation de la qualité esthétique de photographies passe notamment par la différentiation entre le sujet et le fond de la photo. Les méthodes utilisées sont variées. Ainsi, [Wong et Low 2009] emploient des cartes de saillance pour distinguer les zones importantes du reste de l'image. Ces cartes peuvent être obtenues de plusieurs manières. [Wong et Low 2009] exploitent la méthode présentée par [Itti et al. 1998], dont le principe repose sur la fusion de mesures locales de contraste, de contours et de coloration. Ainsi, les zones saillantes sont définies comme

celles contenant de forts contrastes, des contours nets et des couleurs vives. Par la suite, un ensemble de mesures sont calculées, décrivant des aspects tels que des caractéristiques globales de qualité d'image, des informations sur les régions extraites précédemment (taille, nombre et emplacement des zones saillantes) ou encore sur les relations qui peuvent exister entre les zones très saillantes et peu saillantes de la photo (différences de couleurs, de luminosité...). Les résultats des expériences menées par [Wong et Low 2009] montrent qu'à bases de données et algorithmes d'apprentissages identiques, les résultats dépassent d'environ 5% ceux obtenus par [Ke et al. 2006] et [Datta et al. 2006] en terme de bonne classification, et sont très proches de ceux obtenus par [Ng et al. 2009].

Distinction entre avant-plan et arrière-plan

Plusieurs travaux récents utilisent essentiellement des statistiques obtenues à partir des pixels de l'image (moyenne des gradients, écart-type de la luminance) qui permettent d'obtenir des informations pertinentes (netteté, contraste) qui sont à la fois faciles à interpréter et dont le calcul repose sur l'analyse directe des pixels. Un état de l'art des principales statistiques considérées est établi par [Faria et al. 2013]. Des données relatives à la composition de l'image peuvent ensuite être ajoutées : en utilisant la méthode de détection de l'avant-plan proposée par [Luo et Tang 2008], [Faria et al. 2013] calculent séparément des statistiques à la fois sur l'avant-plan et sur l'arrière-plan, afin de mettre en valeur les différences entre ces deux régions. La détection de visages est également considérée car les photos de visage correspondent à des cas particuliers, dont la composition et l'appréciation humaine diffèrent des autres types de photos. Cette particularité est mise en avant dans le travail de [Tang et al. 2013]. Ce travail propose en effet de faire des apprentissages particuliers pour 7 types de photographies différents : animaux, plantes, humains, architecture, paysages, nuit, nature morte. L'idée est d'extraire des informations sur l'image extraites directement au niveau des pixels (contraste, nombre de couleurs, netteté, éclairage, etc.) de l'avant-plan ainsi que de l'arrière-plan, puis d'adapter l'apprentissage supervisé au type d'image considéré. Cela permet également d'adapter le choix des caractéristiques à étudier : dans le cas de photos contenant des personnes, la détection de visages et le calcul de statistiques dans ces régions particulières améliore les performances. Le problème principal de la méthode est le besoin d'une information supplémentaire concernant le type de l'image, ou la nécessité de recourir à une étape de classification préliminaire afin de déterminer le type de l'image.

Différences entre avant-plan et arrière-plan

En plus de calculer des caractéristiques dans chaque région de l'image (avant/arrière-plan), [Kim et Kim 2014] ajoutent des mesures tenant compte des différences entre le sujet et le fond : différence d'éclairage, de netteté, de contraste, de couleurs. Ces mesures, ainsi qu'un algorithme de détection du sujet efficace (cartes de saillance, estimation de netteté, détection de visage), permettent d'améliorer les performances de classification obtenues par [Tang et al. 2013] sur la même base de photos. Les taux moyens de bonne classification sur l'ensemble des 7 catégories de photos présentes sont ainsi augmentés d'environ 3,5% par rapport aux

travaux de [Tang et al. 2013]. Enfin, il est intéressant de noter que de plus en plus de travaux classent les caractéristiques à calculer en différentes catégories. Par exemple, [Aydin et al. 2015] considèrent que la qualité esthétique de photographies s'évalue par la netteté, la profondeur de champ (différence de netteté entre le sujet et le fond de l'image), la clarté (différence de contraste entre le sujet et le fond), l'éclairage moyen et la coloration. Le produit de ces 5 mesures est alors considéré pour donner une évaluation globale de la qualité esthétique. Le principe consistant à distinguer les catégories de mesures présente au moins deux avantages. Le premier est la possibilité d'améliorer les performances globales en combinant des statistiques calculées au niveau du pixel (moyenne des gradients, concentration des gradients dans certaines régions) afin d'obtenir des informations plus proches de la compréhension humaine (impression globale de netteté). Ensuite, ces informations peuvent transmettre des retours sur la qualité perçue de l'image : image trop floue, trop peu colorée, trop peu contrastée, etc. Ces retours sont très intéressants pour la retouche de photos, ou simplement pour aider à leur capture (principe de l'autofocus ou de la balance des blancs dans les appareils photos).

Choix de l'algorithme et des bases d'images

Jusqu'à présent, nous avons essentiellement discuté du choix des caractéristiques à considérer pour prédire la qualité esthétique de photos. Or l'apprentissage automatique supervisé est également une étape cruciale dans l'estimation de la qualité esthétique. De nombreux algorithmes ont été utilisés dans les travaux précédemment cités, dont une liste non exhaustive est donnée en 1.4.2. Il a par exemple été montré par [Datta et al. 2007] que pour la même base de données et le même choix de caractéristiques, une amélioration significative des résultats est possible en combinant les avantages de différents algorithmes (dans ce cas précis, classification bayésienne et machines à vecteurs de support). Toutefois, très peu de travaux comparent l'impact des différents algorithmes d'apprentissage, et lorsque des comparaisons sont effectuées, il ressort généralement que les performances ne varient que très peu lorsque l'algorithme ou ses paramètres sont modifiés. Le choix de la base de données influe également sur la réussite de la prédiction. La difficulté à obtenir une base de données à la fois conséquente (nombreuses images), pertinente (images liées au problème considéré) et correctement évaluée (scores attribués par un nombre significatif de sujets juges) limite souvent les performances des méthodes. En effet, une base contenant trop peu de photos ne permettra pas de concevoir un modèle suffisamment générique. Il en est de même pour un ensemble de photos trop semblables les unes aux autres. Enfin, l'évaluation de ces photos par des humains est subjective et introduit des biais : un grand nombre de votes est donc requis pour avoir une évaluation significative de chaque image. Les différentes bases de données utilisées pour l'étude de la qualité esthétique sont présentées en 1.2.4.

Applications et réalisations

Les travaux précédemment cités ont pour point commun l'étude de la qualité esthétique de photographies. L'intérêt de ces estimations est multiple. Un premier objectif serait de pouvoir trier automatiquement des photos, afin d'en sélectionner certaines pour un partage sur un

réseau social, ou pour créer rapidement un album de photos. Trier les photos permet également d'améliorer la pertinence des moteurs de recherche d'images, en proposant en priorité des images dont la qualité esthétique est élevée ([Ng et al. 2009 ; Luo et Tang 2008]). Des indices de qualité esthétique permettent également de retoucher automatiquement des photographies, comme présenté dans les travaux de [Nishiyama et al. 2009], dont l'objectif est d'optimiser le découpage de la photo de manière à obtenir le recadrage le plus esthétique possible. Des programmes informatiques à visée industrielle ont également vu le jour. Par exemple, [Datta et Wang 2010] ont développé une méthode d'estimation automatique et en temps réel de la qualité esthétique de photographies. Cette méthode est basée sur un algorithme allégé de la méthode présentée par [Datta et al. 2006], et a pour objectif d'aider les photographes amateurs à prendre des clichés de bonne qualité esthétique. Une interface web proposant de trier automatiquement un ensemble d'images selon leur qualité esthétique a été proposée par [Li et al. 2010b]. Cette interface n'est actuellement plus disponible, mais montre une application potentielle de l'estimation automatique de la qualité esthétique de photos de visage. Notons enfin que l'étude de la qualité esthétique ne se limite pas aux photographies et peut être étendue à l'évaluation de peintures (voir les travaux présentés par [Li et Chen 2009]) ou d'autres d'images.

Finalement, le caractère subjectif du problème considéré induit de nombreux problèmes qui seront abordés au fur et à mesure dans ce document. Notamment, [Datta et al. 2008] traitent du décalage qu'il y a entre les estimations que l'on peut faire automatiquement sur une image (calcul de caractéristiques) et les émotions qu'elle dégage : amusement, joie, peur, tristesse, etc. [Joshi et al. 2011] détaillent les différents problèmes rencontrés lors de l'estimation de la qualité esthétique. En effet, le grand nombre d'outils permettant d'effectuer de l'apprentissage, la dépendance des résultats aux bases de données et la difficulté du choix des caractéristiques à considérer sont tous des problèmes ouverts. De plus, le type d'images considéré ainsi que le sujet traité ou les conditions de son visionnement sont des critères importants. Selon l'application choisie (logiciel embarqué dans un appareil photo, logiciel de retouche photo sur ordinateur, traitement en temps réel ou non...) les solutions retenues ne sont pas les mêmes, celles-ci dépendent des capacités de calcul et de mémoire du support, du temps dont dispose l'utilisateur.

1.2.3 Étude de la qualité esthétique de photos de visage

Les études précédentes considèrent essentiellement les photos dans leur ensemble. Or il est rapidement apparu que lorsque l'on cherche à estimer la qualité de photographies, il est intéressant de savoir de quel type de photo il s'agit ([Tang et al. 2013]). Il ne semble pas absurde de penser que les caractéristiques contrôlant la réussite d'une photo de paysage ne s'appliquent pas toutes aux photos de visage : si une photo de paysage est souvent considérée dans son ensemble, l'essentiel de l'attention visuelle est concentrée sur le(s) visage(s) lorsque des personnes se trouvent sur la photo. Il convient alors de définir des critères particuliers adaptés aux photos de visage.

Détection de visages

Ce n'est qu'avec le développement de techniques de détection d'objets rapides et robustes telles que celle décrite dans [Viola et Jones 2001] et de détection du sujet de l'image ([Luo et Tang 2008] ou [Wong et Low 2009]) que des recherches ont été faites dans le cas particulier des photos contenant les visages. Souvent, la technique générale employée est la même que pour les autres types de photos, mais des caractéristiques additionnelles sont calculées dans les zones contenant des visages. Une première idée est présentée dans [Jiang et al. 2010] où un ensemble de caractéristiques utilisées dans des travaux antérieurs (notamment celles introduites par [Ke et al. 2006]) est calculé sur différentes régions de l'image. Ces régions sont les 9 régions délimitées par la règle des tiers, auxquelles sont ajoutées les zones où se situent les visages détectés. Des mesures décrivant la position et la taille des visages sont également prises en compte. Les expériences montrent une amélioration significative du taux de bonne classification lorsque seules les photos de visage sont considérées : les informations apportées par la détection de visage semblent aider à l'estimation de la qualité esthétique.

Photos contenant des visages

Par la suite, [Li et al. 2010a] ne considèrent que des photos contenant des personnes, et s'intéressent aux relations qui peuvent exister entre les visages d'une même photo. Des méthodes automatiques d'extraction de caractéristiques faciales sont implémentées : ouverture des yeux, de la bouche. L'orientation et l'expression du visage sont calculées à l'aide des positions relatives de ces caractéristiques. Ces mesures sont complétées par les positions relatives des différents visages dans la photo : la proximité des visages dans une même photo est généralement appréciée. En mesurant également des critères classiques de qualité esthétique de photographie (couleurs, textures, éclairage...), les expériences effectuées montrent des résultats de classification intéressants. Des tests de régression ont également été tentés afin de prédire des scores de qualité pour chaque image, et les valeurs renvoyées par l'algorithme sont corrélées aux scores de vérité terrain. Les résultats sont sensiblement meilleurs que ceux proposés par l'outil développé dans [Datta et Wang 2010] dans le cas d'images contenant des visages. Un programme de retouche de photos basé sur ces mesures de qualité esthétique a été implémenté et est présenté dans [Li et al. 2010b]. Ce programme est comparé avec des outils grand public tels que le logiciel Picasa [1], qui intègre également des algorithmes permettant la retouche et le recadrage automatique de photographies. Une amélioration des performances obtenues par [Li et al. 2010a] est proposée par [Xue et al. 2013], dont l'idée repose sur la distinction entre les photos contenant un seul visage et celles contenant plusieurs visages. En effet, les mesures telles que les distances entre les visages n'étant pas pertinentes pour des portraits, il semble naturel de considérer des caractéristiques différentes dans les deux cas. Cette distinction permet de diminuer l'erreur de prédiction du score de chaque image de 10% environ.

[Battiato et al. 2013] proposent une évaluation différente dont l'intérêt est de ne pas recourir à la phase d'apprentissage. L'idée est de développer le même type de caractéristiques que celles

1. Voir `http://www.google.com/intl/fr/picasa/`.

utilisées habituellement, mais par la suite, au lieu d'utiliser un algorithme d'apprentissage supervisé, les valeurs calculées sont fusionées à l'aide d'une combinaison linéaire de coefficients. Tout d'abord, les visages sont détectés (nous notons F le nombre de visages détectés), et des indices de couleurs (notés C) relatifs à la saturation sont calculés sur chaque visage. Ensuite, la taille et la position de chaque visage sont prises en compte (notées T et P). À ces critères sont ajoutés des valeurs mesurant l'expression faciale (joie, surprise, colère, tristesse, respectivement j, s, c, t) et la fermeture des yeux (o), qui sont calculées par le logiciel présenté par [Ernst et al. 2009]. Enfin, le score final de l'image correspond à la moyenne des scores de chaque visage, estimés par une combinaison linéaire des coefficients définis précédemment. La formule explicite considérée pour l'évaluation de la qualité esthétique de photos de visage est la suivante :

$$\text{Score} = \frac{1}{F}\sum_{i=1}^{F}[C_i \times (T_i + P_i + 0.1 \times (j_i + s_i - c_i - t_i - o_i))] \tag{1.1}$$

Toutefois peu de comparaisons avec les autres travaux sont effectuées, il est donc difficile d'évaluer la pertinence de la méthode. De plus, l'estimation des caractéristiques liées aux expressions faciales est également un problème difficile, et il n'est pas certain que ces paramètres soient toujours fiables. Enfin, la méthode tient uniquement compte des régions correspondant aux visages qui peuvent être petites par rapport à la photo entière, ce qui ne suffit pas toujours à apprécier l'ensemble de l'esthétisme de l'image.

Portraits

[Khan et Vogel 2012] sont parmi les premiers à restreindre l'étude de la qualité esthétique au cas des portraits. Ils définissent des caractéristiques particulières relatives aux effets photographiques présents dans les images de portrait : différence de luminosité entre la partie gauche et la partie droite du visage, répartition de l'éclairage, contraste entre l'avant-plan et l'arrière-plan du visage. La base de photos utilisée est la même que celle élaborée par [Li et al. 2010a], mais [Khan et Vogel 2012] ne conservent que les photos contenant un seul visage, soit 145 photos. En comparant leurs performances avec celles des travaux antérieurs qui n'utilisent pas de caractéristiques spécifiques aux visages, le taux de bonne classification passe de 58 à 61% dans le cas d'études sur deux catégories. En plus de cela, seules 7 mesures sont effectuées sur la photo, contre 63 pour [Li et al. 2010a], ce qui montre la grande pertinence de ces caractéristiques dans le cas de photos de visage. Un autre intérêt de cette méthode est la possibilité de l'appliquer à d'autres types de photos (photos de voitures, d'animaux...), la seule limite étant la capacité à détecter les objets en question. Il est important de noter que l'étude n'a été réalisée que sur un ensemble de 145 photos, ce qui ne suffit pas pour généraliser les performances de la méthode à l'ensemble des photos de portraits : il est possible que les performances soient plus faibles sur des bases d'images plus grandes.

[Pogačnik et al. 2012] implémentent 71 caractéristiques variées (toutes ne sont pas détaillées) décrivant différents aspects de la photo. Toutefois, des caractéristiques propres aux photos de visage sont considérées, et une base de 1048 portraits est constituée à partir du site [*DPChallenge*]. Ces photos sont déjà évaluées, chacune par une centaine d'utilisateurs du site,

selon leur qualité esthétique sur une échelle de 1 à 10. Une analyse des caractéristiques pertinentes est effectuée à l'aide de l'algorithme *ReliefF*, dont le principe est expliqué en détail dans [Robnik-Šikonja et Kononenko 2003]. Cet algorithme repose sur l'hypothèse que pour deux images de qualité esthétique différente, une caractéristique discriminante prendra des valeurs différentes. Dans leur analyse, il est montré qu'une photo de bonne qualité esthétique présente un nombre réduit de teintes différentes. Le sujet ne doit être ni trop petit ni trop grand par rapport à la taille de la photo, et ne doit pas être situé sur les bords de l'image. Enfin, il ressort de l'étude que la plupart des contours, et la proportion des contours dans la région du sujet jouent un rôle important dans la perception de la qualité esthétique d'un portrait. L'algorithme d'analyse des caractéristiques présente ainsi un double avantage. En effet, cela permet d'obtenir des informations concernant les critères les plus pertinents. De plus, la connaissance des caractéristiques les moins discriminantes permet de les écarter du modèle, afin d'en améliorer les performances ou de réduire le temps de calcul. Les résultats présentés dans ces travaux confirment cela, puisqu'en conservant seulement 41 des 71 caractéristiques, les taux de bonne classification sur leur base de portraits passent de 73,4 à 74,8%.

Photos de visage

Nous venons de voir deux exemples de travaux concernant les portraits. Dans ce document, nous étudions plus particulièrement les photos de visage, que nous définissons comme étant des photos découpées au niveau des épaules, dont le visage est suffisamment grand par rapport à l'image entière. Souvent, le visage est centré, présenté de face et ne présente pas d'orientation particulière. La différence principale entre les photos de visage et les portraits est la taille du visage dans la photo. Un premier travail sur ce type d'images est présenté par [Males et al. 2013]. 10 caractéristiques sont calculées : une mesure de netteté basée sur le gradient de l'image, la profondeur de champ qui traduit la différence de netteté entre l'avant-plan et l'arrière-plan, trois mesures de distance entre le centre du visage et les points définis par la règle des tiers (indice de composition), la moyenne et l'écart type de la luminance qui attestent respectivement de la luminosité et du contraste, le nombre de pixels dont l'intensité est maximale (indice de surexposition), le nombre de teintes différentes représentées dans l'image (indice de coloration), et le rapport entre la taille du visage détecté et la taille de l'image, ce qui permet de s'assurer que le visage est clairement le sujet principal de la photographie. Les expériences sont effectuées sur une base d'images que les auteurs ont constituée eux-mêmes et peu de comparaisons avec les autres travaux existants sont évoquées. Les résultats présentés montrent que ces caractéristiques sont suffisantes pour obtenir de très bons résultats de classification sur la base d'images considérée, et ces conclusions sont cohérentes avec celles données par [Khan et Vogel 2012] et [Pogačnik et al. 2012] : peu de caractéristiques pertinentes sont suffisantes, et certaines d'entre elles doivent encoder des informations relatives à la mise en valeur du visage (profondeur de champ, taille et position du visage, etc.).

Actuellement

En 2015, [Redi et al. 2015] réalisent de nombreuses expériences sur les portraits contenant

des visages, à partir de plus de 10000 photos extraites du site [*DPChallenge*]. Des caractéristiques très diverses sont extraites et étudiées afin de formuler des modèles d'estimation de qualité esthétique plus performants que ceux présentés dans l'état de l'art. Les auteurs considèrent notamment divers attributs du visage, du même type que ceux extraits par [Battiato et al. 2013] : présence de sourire, de lunettes, mais également âge, sexe, position et orientation du visage, couleur de la peau, position des yeux, du nez et de la bouche. Ces attributs sont déterminés à l'aide de l'outil [*FacePlusPlus*]. Cet outil prend en entrée une photo de visage et renvoie automatiquement la liste des attributs. Le détail de l'implémentation du calcul de ces attributs n'est pas disponible, il est donc difficile d'évaluer la fiabilité de ces mesures. De plus il est montré [Redi et al. 2015] qu'en dehors du sourire, la corrélation de ces attributs avec la qualité esthétique des images est faible ou nulle. D'autres mesures concernant la qualité de l'image sont étudiées : taux de compression JPEG, estimation du bruit, etc. Enfin, de nombreuses caractéristiques décrivant la texture, le contraste et la luminosité sont calculées. Il ressort des expériences menées sur les 10000 portraits que les caractéristiques les plus pertinentes sont les mesures de netteté dans les régions du visage correspondant aux yeux, ce qui sera confirmé et détaillé dans le présent document. En outre, les résultats sont significativement supérieurs à ceux rapportés par [Khan et Vogel 2012] : les performances de classification sur une même base de données passent de 63 à 74%.

Conclusion

Finalement, l'approche générale permettant d'estimer la qualité esthétique de photographies est toujours sensiblement la même et repose sur une extraction de caractéristiques puis sur un apprentissage supervisé à l'aide d'une base de données annotée par des humains. Le travail présenté dans ce document s'inscrit dans la continuité de cette approche. Les principales améliorations apportées par les travaux sur l'estimation de la qualité esthétique d'une photo de visage sont la prise en compte de la composition de l'image, puis de l'avant-plan, du visage et enfin de caractéristiques à l'intérieur du visage (sourire, netteté des yeux [Redi et al. 2015]). Les bases de photos considérées ont également évolué, pour passer de quelques centaines de photos étudiées par des volontaires à d'importantes bases de données collectées sur des sites de partage de photos. Une liste non exhaustive des principales bases de données existantes est donnée dans la section suivante.

1.2.4 Principales bases de photos évaluées

Dans cette section sont présentées les différentes bases de photos construites et utilisées dans le cadre de l'évaluation de la qualité esthétique de photographies de visages. Les bases présentées contiennent au moins plusieurs centaines de photos afin d'avoir un nombre significatif d'images pour l'apprentissage et les tests, et sont annotées par des humains selon leur qualité esthétique. Ces annotations peuvent être des scores correspondant à la moyenne des scores donnés par plusieurs humains (pour les problèmes de régression), ou des labels définissant la catégorie à laquelle correspond l'image : photo de bonne ou de mauvaise qualité esthétique (pour les problèmes de classification).

Nous ne présentons ici que brièvement les différentes bases considérées dans l'état de l'art.

1.2.4.1 Bases de photos collectées sur internet

Récupérer automatiquement des photos sur Internet permet d'obtenir les bases de photos les plus conséquentes, de plusieurs milliers à plusieurs millions de photos. De nombreux biais sont toutefois introduits dans les évaluations de ces photographies. Les personnes évaluant les photographies n'ont en effet pas toujours de consigne claire, la qualité esthétique étant une notion subjective : doit-on évaluer la photo dans son ensemble ? Tenir compte du sujet, de son originalité ? De plus les internautes auront tendance à privilégier les photos de leurs amis ou contacts, certaines photos auront tendance à être moins notées du fait de leur impopularité. De manière générale, les photos publiées sont plutôt réussies ; l'échelle de notation n'est pas entièrement utilisée car les photographes auront plus facilement tendance à partager leurs réussites que leurs échecs. L'apprentissage est ainsi facilité par le grand nombre d'échantillons disponibles, mais limité par la faible fiabilité des votes et la faible répartition des scores disponibles. 3 principaux sites sont généralement utilisés pour étudier la qualité esthétique de photographies.

Photo.net

[*PhotoNet*] est un site de partage de photographies. Les images sont classées selon les notes attribuées par au moins 5 utilisateurs selon leur qualité esthétique. Les notes s'étalent sur une échelle de 1 à 7, et la distribution des scores pour chaque photo n'est pas disponible. Différents travaux basent leurs expériences sur des photos récupérées sur [*PhotoNet*] : [Datta et al. 2006 ; Wong et Low 2009 ; Marchesotti et Perronnin 2011]. [Datta et al. 2006] ont récupéré automatiquement un ensemble de 3500 photos qu'il est possible de télécharger et qui sert de point de comparaison pour les travaux ultérieurs. Les performances actuelles de classification et de régression sur cette base sont résumées dans le tableau 1.1. Les résultats rapportés sont obtenus en ne considérant que 20% des images dont les scores sont les plus extrêmes afin de faciliter la classification. En ajoutant les photos de score intermédiaire, les performances chutent d'environ 10% pour chacun des travaux considérés. Les photos issues de [*PhotoNet*] et utilisées par les travaux cités ne sont pas forcément des photos de visage (paysages, animaux, etc.).

Flickr

[*Flickr*] est un site de partage de photographies, utilisé par les amateurs et les professionnels. Ce site propose un moteur de recherche créé par *Yahoo !* et est basé sur l'intérêt que portent les utilisateurs aux photos. La méthode employée n'est pas dévoilée, mais les critères peuvent être l'origine du cliché, les commentaires, les mots clés associés à la photo, le nombre de partages de la photo. Si [*Flickr*] met à disposition un grand nombre d'images ainsi que la possibilité de trier ces images par ordre de pertinence (selon les critères du moteur de recherche), cela ne permet pas d'en évaluer directement la qualité esthétique. Le recours à des

TABLEAU 1.1 – Performances de différents travaux dont les évaluations sont faites sur des images récupérées sur [*PhotoNet*]. Le Taux de Bonne Classification (T_{BC}) et l'Erreur Quadratique Moyenne (EQM) permettent de quantifier ces performances.

	Nombre d'images considérées	Performances de classification (T_{BC})	Performances de régression (EQM)
[Datta et al. 2006]	3581	70%	0,5
[Wong et Low 2009]	3161	79%	/
[Marchesotti et Perronnin 2011]	3118	78%	/

TABLEAU 1.2 – Performances de différents travaux dont les évaluations sont faites sur des images récupérées sur [*Flickr*]. Le Taux de Bonne Classification (T_{BC}) et l'Erreur Quadratique Moyenne (EQM) permettent de quantifier ces performances. Les travaux comparables entre eux (car se basant sur les mêmes photos) sont regroupés dans une même cellule du tableau.

	Nombre d'images considérées	Performances de classification (T_{BC})	Performances de régression (EQM)
[Pogačnik et al. 2012]	114	95%	/
[Males et al. 2013]	380	86%	/
[Li et al. 2010a]	500	68%	2,38
[Xue et al. 2013]	500	/	2,11
[Khan et Vogel 2012]	145	63%	/
[Redi et al. 2015]	145	74%	/

évaluations par des humains est donc toujours nécessaire, ce qui limite les tailles des bases de données provenant de ce site. Plusieurs articles font référence à l'utilisation des photos partagées sur [*Flickr*]. C'est le cas par exemple pour [Cerosaletti et Loui 2009 ; Jiang et al. 2010 ; Pogačnik et al. 2012], qui se limitent par conséquent à une sélection de quelques centaines de photos, évaluées ensuite par des humains.

Un bref résumé des résultats obtenus sur des bases d'images issues de [*Flickr*] est donné dans le tableau 1.2. Les résultats présentés dans ce tableau sont obtenus sur des photos contenant des visages.

DPChallenge

[*DPChallenge*] est également un site regroupant des photographies d'amateurs et de professionnels. Des concours y sont proposés, où l'objectif est de présenter la photo la plus réussie correspondant à un thème donné. L'intérêt de ce site est la possibilité pour les participants de noter et de commenter la qualité esthétique des photos sur une échelle de 1 à 10. De nombreux travaux ([Ke et al. 2006 ; Luo et Tang 2008 ; Ng et al. 2009 ; Desnoyer et Wettergreen 2010 ; Dhar et al. 2011 ; Pogačnik et al. 2012 ; Tang et al. 2013 ; Kim et Kim 2014 ; Redi et al. 2015]) se sont appuyés sur ce site, en téléchargeant automatiquement un grand nombre de photos et

TABLEAU 1.3 – Performances de différents travaux dont les évaluations sont faites sur des images récupérées sur [*DPChallenge*]. Le Taux de Bonne Classification (T_{BC}) permet de quantifier ces performances. Les travaux comparables entre eux (car se basant sur les mêmes photos) sont regroupés dans une même cellule du tableau.

	Nombre d'images considérées	Performances de classification (T_{BC})
[Ke et al. 2006]	6000	72%
[Ng et al. 2009]	2000	80%
[Dhar et al. 2011]	3200	80%
[Desnoyer et Wettergreen 2010]	4955	63%
[Pogačnik et al. 2012]	1048	75%
[Tang et al. 2013]	17613	90%
[Kim et Kim 2014]	17613	92%
[Redi et al. 2015]	10141	64%

les scores qui leur sont associés. Ainsi Ke et al. ont construit un ensemble de 60000 photos. Un autre ensemble de 250000 photos triées par thématique est présenté et étudié par Murray et al. Enfin, Tang et al. considèrent 17613 photos réparties en 7 catégories (paysages, portraits, etc.) et regroupées en deux classes : les photos de très bonne et de très mauvaise qualité esthétique. Ces photos sont réutilisées dans les travaux de [Kim et Kim 2014] [2]. [*DPChallenge*] permet donc de fournir un grand nombre de photos déjà évaluées par un nombre significatif de sujets (au moins 78 dans la base de [Murray et al. 2012]). De plus la distribution des scores est fournie pour chaque image ce qui permet d'avoir une idée du consensus entre les votants. Un bref résumé des résultats obtenus à partir de ces images est donné dans le tableau 1.3. Il est important de noter que dans ce tableau, les performances au-delà de 70% de bonne classification (Ke et al. ; Ng et al. ; Pogačnik et al. ; Tang et al. ; Kim et Kim) correspondent à des évaluations où seules des photos aux scores extrêmes sont considérées. De manière générale, le site [*DPChallenge*] contient des photographies plutôt réussies, et l'écart-type des scores des photos est très faible. Enfin, seuls les travaux de Pogačnik et al. ; Tang et al. ; Kim et Kim ; Redi et al. considèrent des photographies contenant des visages.

1.2.4.2 Autres bases de photos

Les bases de photos qui n'ont pas été collectées sur internet sont généralement bien moins fournies : en général seules quelques centaines d'images sont considérées. Parfois, les bases de photos ne sont pas annotées. Par exemple, pour tester leur méthode de classification, [Tong et al. 2004] utilisent deux bases de données distinctes, représentant deux types d'images différentes : des images récupérées sur des sites de photographes professionnels sont automatiquement considérées comme étant de bonne qualité esthétique, tandis que des images provenant de sites regroupant des photographes amateurs sont considérées comme de moins bonne qualité esthétique. Il existe par ailleurs différentes bases conséquentes de photos de visage, mais

2. Photos disponibles en téléchargement à l'adresse `http://mmlab.ie.cuhk.edu.hk/CUHKPQ/Dataset.htm`.

ces dernières sont très souvent utilisées dans le cadre de la reconnaissance faciale, et ne sont donc pas évaluées en terme de qualité esthétique. De plus les visages ne sont pas toujours présentés de face.

Un exemple de base de données constituée de photos sélectionnées manuellement, puis évaluées à l'aide de volontaires, est donné dans [Cerosaletti et Loui 2009]. Ce jeu de données a été réutilisé par la suite dans [Jiang et al. 2010]. Au total, 450 images sont choisies à la fois en fonction du type de sujet et de la taille du sujet par rapport à la photo dans son ensemble. Des photos ont été prises à l'intérieur, d'autres en plein air ; certaines présentent des humains ou animaux et d'autres des objets. L'origine des photos peut être variée. Celles présentées ici proviennent du site de partage [*Flickr*], du *Kodak Picture of the Day* ou encore de collections privées. Une trentaine de participants a pour tâche de classer les photos, l'objectif final étant d'obtenir des notes sur une échelle discrète allant de 1 à 100. Ce score est finalement défini comme étant la vérité terrain, utilisée pour les expériences. Beaucoup de bases de données suivent ce modèle, malheureusement celles-ci sont rarement disponibles.

Finalement, très peu de bases de données utilisées et présentées sont directement réutilisables ; la plupart étant le résultat d'études qualitatives et quantitatives internes, elles ne sont que rarement publiées avec les articles. Nous décrivons les bases de données que nous avons utilisées dans notre travail dans le chapitre 3.

1.3 Impressions induites par une photo de visage

1.3.1 Importance des premières impressions et du contexte

À partir d'une photo de visage, il est possible de se faire une opinion à propos de l'identité d'une personne, de son état émotionnel (joie, tristesse, colère, etc.), de ses intentions (visage amical ou menaçant), de son état de santé (teint de la peau, traits du visage), de son attractivité, de son âge, etc. Ainsi, lorsqu'un individu voit une personne inconnue sur une photo, il lui infère d'emblée des traits de personnalité, essentiellement à partir des caractéristiques du visage. Les recherches menées en sciences sociales ont montré que chacun se fait rapidement et automatiquement une première impression sur un individu à partir d'une simple photo, et que ce premier jugement est difficile à modifier par la suite. [Willis et Todorov 2006] réalisent une expérience afin de montrer que 100 millisecondes suffisent pour qu'un jugement soit donné. Dans ces travaux, il est également montré que ce jugement ne varie que très peu lorsque la durée de visualisation augmente, et le jugement ne change que rarement lorsqu'aucune limite de temps n'est donnée. Ces évaluations sont faites avec un degré élevé de consensus entre les individus, et concernent des caractéristiques diverses telles que la compétence, l'amabilité, l'honnêteté, la fiabilité ou la sympathie. Ces expériences montrent que malgré le caractère subjectif de ces évaluations, il existe des éléments objectifs que les humains ont tendance à apprécier de la même manière : un sourire est par exemple interprété comme un signe de sympathie.

Il ressort de ces constats que le choix d'une photo de visage particulière est crucial lorsque celle-ci a pour objectif de présenter un individu : photo de profil, illustration d'un CV, etc. Ce choix est d'autant plus important qu'il peut avoir des conséquences : une photo dont le visage est jugé attractif aura plus de succès sur un site de rencontre, un visage qui dégage une impression de confiance, d'honnêteté et de compétence aura un impact significatif sur une affiche de campagne électorale. Il est important de garder à l'esprit que nous évaluons dans ces travaux l'impression dégagée par une image, et en aucun cas nous ne cherchons à inférer la qualité intrinsèque d'un individu. Nous ne cherchons pas à évaluer les traits communs aux personnes compétentes, mais les traits communs aux photos de personnes évaluées comme étant compétentes. Bref, nous essayons dans ces travaux de reproduire les évaluations subjectives faites par des humains, l'objectif étant de créer un algorithme présentant les mêmes faiblesses que les évaluations humaines (présence de clichés, de préjugés, etc.).

Ce travail se rapproche des recherches dans le domaine de l'évaluation automatique des expressions faciales et des émotions. En effet, les traits du visage qui définissent ces expressions sont les premiers critères considérés lors de l'étude d'une photo de visage : la personne sourit-elle ? Semble-t-elle joyeuse ou plutôt triste ? De nombreuses méthodes permettant d'estimer les émotions existent, et ne sont pas traitées dans ce document. Nous pensons que le problème qui se pose ici est légèrement différent, et que l'analyse des expressions et des émotions n'est qu'une partie de notre problème, qui traite de notions encore plus abstraites et subjectives telles que les impressions de compétence et de sympathie dégagées par une photo. La reconnaissance d'émotions ne repose que sur les traits du visage, alors que notre problème tient compte de la photo dans son ensemble : orientation et position du visage, vêtements, bijoux, fond de la photo, etc.

Avant de s'intéresser aux modèles d'évaluation, il est important de s'intéresser aux différents traits de caractère qui sont évalués lorsqu'une personne considère une photo. Dans leurs expériences, [Oosterhof et Todorov 2008] proposent aux participants de noter tout ce qui leur vient à l'esprit par rapport aux visages représentés sur 66 photos. Les évaluations fournies par les participants représentent des informations sur le caractère (agressif, aimable), l'apparence (mal habillé, mal coiffé), sur les habitudes (sportif), etc. Ces évaluations sont alors regroupées par les chercheurs, qui définissent ainsi 14 traits principaux : attractif, attentionné, agressif, méchant, intelligent, assuré, émotif, digne de confiance, responsable, sociable, étrange, triste, dominant, menaçant. Par la suite, [Todorov et al. 2008] montrent que ces traits peuvent être réduits à deux principales évaluations, en effectuant une Analyse en Composantes Principales (ACP). Par cette analyse, il est également montré que les deux traits "digne de confiance" et "dominant" sont fortement corrélés aux autres évaluations et chacun des autres traits peut s'exprimer comme une combinaison linéaire de ces caractéristiques.

1.3.2 Méthodes d'estimation des traits de caractère

Une première méthode permettant de proposer une évaluation automatique des 9 différents traits de caractère suivants est discutée dans les travaux de [Todorov et Oosterhof 2011] : dominant, menaçant, attractif, effrayant, méchant, digne de confiance, extraverti, compétent,

sympathique. Dans ces travaux, 300 visages sont générés automatiquement par le logiciel Facegen Modeller [3] avec différents paramètres de forme du visage et de couleur de peau. Les visages sont tronqués au niveau du cou, ne possèdent ni cheveux, ni accessoires, et l'arrière-plan de l'image est entièrement noir. Ainsi, seuls les paramètres de forme de visage et de réflectance (intensité, couleur et texture de la peau) sont pris en compte. Ces paramètres, au nombre de 2043 pour la forme (grille définissant le visage en $3D$) et $3 * 256 * 256$ pour la réflectance (nombre de pixels sur le visage), sont fixés lors de la création du visage par le logiciel. Les paramètres des 300 visages générés sont ensuite réduits par ACP pour que chaque image puisse être décrite par un ensemble de 50 paramètres de forme et 50 paramètres de réflectance. Une expérience est alors réalisée, où les participants ont pour objectif d'évaluer les visages selon chacun des 9 traits. Les paramètres du visage sont alors adaptés aux scores fournis par les participants afin de construire un modèle de régression pour chaque type de jugement. Ces modèles peuvent enfin être considérés afin de créer de nouveaux visages aux traits choisis.

D'autres expériences d'estimation automatique des traits de caractère sont proposées par [Rojas et al. 2011], où les auteurs comparent les performances de classification obtenues par l'utilisation des positions des points de repère à l'utilisation de descripteurs génériques *HOG* [Dalal et Triggs 2005]. Il en ressort que les descripteurs génériques sont légèrement plus performants que les points de repère, même lorsque ces derniers sont combinés de manière à intégrer des relations entre les points : angles formés par 3 points adjacents, distances entre les points. Rojas et al. basent leurs expériences sur les descripteurs *HOG*, mais d'autres descripteurs peuvent également être utilisés afin de décrire les variations d'intensité locales, notamment les filtres de Gabor, très largement utilisés dans le cadre de la reconnaissance faciale [Lajevardi et Lech 2008]. Si les performances ne sont pas similaires pour chacun des traits étudiés, il est montré que dans tous les cas, les algorithmes sont significativement plus performants que le hasard.

Les modèles définis par [Todorov et Oosterhof 2011] sont utilisés dans les travaux de [Todorov et al. 2013] dans le but de générer des bases de données synthétiques de visage correspondant à un certain niveau d'expression d'un trait particulier : visage très menaçant, peu sympathique, très peu digne de confiance, moyennement extraverti, etc. Ces bases, au nombre de 7, sont détaillées en section 1.3.3 et dans le chapitre 4. Des expériences subjectives sont alors réalisées pour valider les modèles d'évaluation définis par [Todorov et Oosterhof 2011] : pour chaque image, des participants tentent de retrouver le niveau d'expression du trait considéré. Les expériences montrent une corrélation importante entre les jugements humains et les modèles de visage, ce qui confirme la pertinence des modèles définis par [Todorov et Oosterhof 2011]. Ces modèles permettent donc de proposer une évaluation objective des traits de personnalité suggérés par une photo de visage.

Comme nous l'avons évoqué dans l'introduction, nous nous limitons dans ce travail à l'évaluation de la sympathie et de la compétence. Nous montrons que l'impression de sympathie dégagée est très largement liée à la présence de sourire ou d'une expression joviale. L'impression de compétence est souvent augmentée par la présence d'un air sérieux (expression neutre, yeux

3. Voir `http://www.facegen.com`.

ouverts), d'un arrière-plan simple, et d'autres éléments qui tendraient à montrer le savoir et le savoir-faire d'une personne. L'impression de compétence peut ainsi être interprétée comme la confiance que peut avoir un recruteur envers un candidat potentiel.

Les modèles définis ci-dessus incluent des informations sur la forme du visage, les expressions faciales (encodées par les positions des points de repère autour des différents éléments du visage, notamment les sourcils, les yeux et la bouche), ainsi que des informations relatives aux textures, à la couleur et à l'illumination du visage. La principale limite concernant ces modèles, créés à partir de visages synthétiques, est l'absence de nombreuses informations pertinentes dans le cas d'une photo réelle (vêtements, arrière-plan, accessoires). Actuellement, il n'existe pas à notre connaissance de modèle intégrant toutes ces informations. Dans le travail présenté dans ce document, nous tentons de tenir compte de certains de ces éléments, et donc d'améliorer les modèles existants en incorporant des informations sémantiques dans les modèles d'évaluation. L'approche considérée dans ce document passe ainsi par l'utilisation d'attributs correspondant à des informations proches du jugement humain et difficilement calculables à partir des pixels : présence de sourire, de lunettes, état émotionnel, position des sourcils, etc.

Nous pensons que l'utilisation de ces informations permet d'obtenir des estimations plus fiables des impressions de sympathie ou de compétence dégagées par une photo de visage. La limite de cette approche, déjà évoquée dans la section 1.2.3, est la capacité des algorithmes existants à produire une estimation correcte de ces attributs. L'évaluation de la compétence ou de la sympathie est donc limitée par la précision des algorithmes de détection de sourire, de lunettes, etc. Actuellement, plusieurs méthodes existent, et celles-ci reposent sur des algorithmes de détection de visage, d'attributs faciaux (yeux, bouche), ainsi que sur des algorithmes estimant les positions de points de repère dans le visage (limites de la bouche, des yeux, du nez, des sourcils, etc.). La position de ces points de repère peut alors donner des informations sur l'état émotionnel (orientation des sourcils, ouverture de la bouche, etc.), tandis que l'analyse des pixels au niveau des yeux permet par exemple de deviner la présence ou l'absence de lunettes. De nombreuses méthodes d'estimation de chaque attribut sont présentes dans l'état de l'art, toutefois il est particulièrement difficile de produire une implémentation performante pour chacun de ces attributs, c'est pourquoi nous nous limitons aux résultats fournis par les 3 outils d'analyse faciale que sont [*Betaface*], [*SkyBiometry*] et [*SHORE*]. Plus de détails concernant les informations obtenues par ces outils sont donnés dans chapitre 4 de ce document, dans lequel nous présentons également leurs avantages et leurs inconvénients.

1.3.3 Principales bases de visages annotées

Comme dans le cas de l'estimation de la qualité esthétique, nous ne présentons ici que brièvement les bases considérées dans l'état de l'art. Plus de détails ainsi que des exemples de photos et de scores pour chacune des bases utilisées dans ce travail sont donnés dans le chapitre 4, où nous présentons également une base construite dans le cadre de ce travail.

Il existe très peu de bases d'images contenant des visages évalués selon les impressions de

compétence et de sympathie, et nous retenons essentiellement les bases synthétiques proposées par [Todorov et Oosterhof 2011 ; Todorov et al. 2013]. Une première base contient 300 visages générés par [*FaceGen*] et annotés selon les 9 traits suivants : dominant, menaçant, attractif, effrayant, méchant, digne de confiance, extraverti, compétent, sympathique. Les annotations correspondent aux évaluations moyennes fournies par plus de 20 participants.

Par la suite, [Todorov et al. 2013] présentent 7 bases obtenues en exploitant les annotations des 300 visages de la première base. Ces 7 bases correspondent à chacun des 7 traits suivants : attractif, compétent, dominant, extraverti, aimable, menaçant, digne de confiance. Pour chaque base, 25 visages distincts sont générés aléatoirement par le logiciel FaceGen. Ensuite, les caractéristiques du visage sont exagérées (déformation du visage, changement de la couleur de la peau) afin d'augmenter ou d'atténuer le trait de caractère, en fonction des modèles de forme et de réflectance appris en effectuant une régression sur la base de 300 visages annotés. Chaque visage est ainsi modifié sur 7 niveaux d'expression de ce trait, par exemple de très antipathique à très sympathique pour le trait "sympathique". 7 bases de 7×25 visages sont donc créées.

Nous n'avons trouvé qu'une seule base contenant des visages réels et annotés selon différents traits de caractère. Celle-ci est décrite dans les travaux présentés par [Oosterhof et Todorov 2008], dans lesquels 66 visages d'hommes et de femmes sont évalués selon des traits de caractère variés : menace, confiance, dominance, etc. Toutefois, il est difficile de construire des modèles d'évaluation des traits de caractère pertinents à partir d'une base aussi réduite, c'est pourquoi nous avons construit une base de photos plus conséquente qui est détaillée au chapitre 4.

Finalement, le problème de l'évaluation automatique des traits de caractère à partir de photos de visage n'a pas encore été largement traité : de nombreux éléments influant sur les impressions dégagées par les photos ne sont pas pris en compte dans les modèles. Ce problème est accentué par le manque de bases de données conséquentes et variées. Dans ce document nous nous limitons à l'évaluation de la compétence et de la sympathie.

L'estimation de la qualité esthétique et des impressions de compétence et de sympathie véhiculées par une photographie de visage passe par l'extraction de caractéristiques discriminantes. Ces caractéristiques sont extraites dans un premier temps sur des bases de photographies annotées, puis un algorithme d'apprentissage supervisé permet d'associer les caractéristiques aux annotations. Un bilan des caractéristiques ainsi que des algorithmes d'apprentissages de l'état de l'art est donné dans la section 1.4.

1.4 Résumé des caractéristiques et des algorithmes d'apprentissage

1.4.1 Caractéristiques considérées

Dans nos travaux, nous ne considérons que les informations pouvant être extraites à partir des pixels de l'image, ce qui exclut le matériel, les conditions de capture et de visualisation des photographies. Nous faisons la distinction entre trois types de caractéristiques.

Le premier type de caractéristiques correspond aux mesures qui s'effectuent directement à l'aide de l'étude des pixels, qui peuvent traduire des mesures locales (par exemple un dictionnaire des descripteurs SIFT calculés sur l'image) ou des statistiques globales. Ces statistiques correspondent par exemple à des histogrammes de couleurs, aux coefficients de transformées de Fourier, à des valeurs de contraste, de luminosité moyenne, etc. Une liste non exhaustive des caractéristiques de ce type est donnée en 1.4.1.1. Par la suite, nous les considérons comme étant des mesures "de bas niveau" car leur calcul ne nécessite aucune connaissance a priori sur l'image, notamment concernant sa composition et son sujet.

Il est également possible d'extraire les informations "de haut niveau" présentées en 1.4.1.3, qui décrivent des notions plus proches de la perception humaine. Ces caractéristiques correspondent à des attributs qui donnent des informations sémantiques concernant les différents éléments de l'image. Elles requièrent des techniques d'analyse d'image élaborées : extraction du sujet de l'image, détection de contours (avant/arrière-plan), apprentissage de métriques permettant l'évaluation de certains critères (détection de sourires, de fermeture des yeux). Du fait de leur proximité avec le jugement humain, les mesures de haut niveau produisent généralement de meilleurs résultats, mais sont plus difficiles à implémenter et ne sont pas toujours fiables à cause des limites des différents algorithmes de détection, de segmentation ou de classification.

Il est à noter que toutes les caractéristiques de bas niveau peuvent être mesurées dans des sous-régions d'une image. En effet, il peut être intéressant de constater les différences de contraste, de couleur ou de textures entre l'arrière-plan d'une image et son sujet, notamment dans le cas de photos de visage. Ces mesures, qui correspondent à l'analyse directe des pixels (donc de bas niveau) mais nécessitent des informations complexes sur l'image telles que la position d'un visage (information de haut niveau) sont, dans ce document, décrites comme étant des mesures de niveau intermédiaire (voir 1.4.1.2). Ces mesures correspondent essentiellement à la composition des images.

Concernant l'estimation des impressions de compétence et de sympathie, les méthodes actuellement utilisées se basent sur l'extraction de descripteurs génériques (*HOG* ou filtres de Gabor), et les positions de points de repère dans le visage. Ici également, nous nous limitons aux informations contenues dans les pixels. Nous ne cherchons par exemple pas à obtenir d'informations sur l'identité de la personne. Il est évident que si la personne sur la photographie est connue de l'observateur (célébrité, connaissance), les impressions de compétence et de sympathie dégagées seront corrélées à cette information.

1.4.1.1 Mesures de bas niveau

Dans le tableau 1.4 sont résumées les différentes mesures effectuées directement à partir des valeurs des pixels d'une image. La première colonne dénomme la caractéristique, la seconde donne des informations sur la manière de la calculer. La dernière colonne présente les différents travaux ayant utilisé ce descripteur dans leurs expériences d'évaluation de la qualité esthétique ou des impressions de sympathie et de compétence.

Nous avons choisi de classer ces caractéristiques dans 5 catégories : les descripteurs génériques, les indices de texture et de netteté, les indices d'illumination, de contraste, et les mesures de couleurs. Hormis les descripteurs génériques, nous pensons que ces catégories correspondent à la façon dont les humains évaluent l'esthétisme d'une image : est ce qu'elle est nette ? Bien éclairée, contrastée ? Les couleurs sont-elles bien choisies ?

Tableau 1.4 – Caractéristiques de bas niveau utilisées dans l'état de l'art.

Caractéristique	Méthode de calcul / Implémentation	Utilisé par...
Descripteurs génériques		
GIST (signature)	Fusion d'histogrammes de couleur, de contours et d'illumination, réduits par PCA. Implémentation détaillée dans [Oliva et Torralba 2001 ; Siagian et Itti 2007]	Desnoyer et Wettergreen ; Marchesotti et Perronnin ; Xue et al.
Sacs de mots visuels	Dictionnaire de descripteurs SIFT ou SURF	Marchesotti et Perronnin ; Riaz et al.
Histogramme des gradient orientés (HOG)	Histogrammes des gradients orientés calculés dans différents blocs de l'image.	Rojas et al.
Filtres de Gabor	Convolution de l'image par des filtres de Gabor selon différentes orientations à différentes échelles.	Lajevardi et Lech
Netteté et Textures		
Énergie	Moyenne des gradients de l'image	Khan et Vogel ; Kim et Kim ; Redi et al.
Intensité Spectrale et Spatiale	Moyenne géométrique des variations totales (spatiales) et des magnitudes (spectrales) locales	Males et al.
Flou	Différences de variations entre l'image originale et l'image floutée	Crete et al.
Aire des contours	Surface contenant 90% des contours	Ke et al. ; Desnoyer et Wettergreen ; Jiang et al.

Caractéristique	Méthode de calcul / Implémentation	Utilisé par...
Histogrammes	Histogrammes obtenus à partir des filtres de Sobel, Canny ou Laplace	Tong et al. ; Kim et Kim
Transformées	Transformée de Fourier	Ke et al. ; Ng et al. ; Wong et Low ; Desnoyer et Wettergreen ; Xue et al.
	Transformée en Cosinus	Tong et al.
	Transformée en Ondelettes	Tong et al. ; Datta et al. ; Wong et Low ; Riaz et al.
Matrice de Co-Occurence	Énergie, contraste, entropie et homogénéité de la matrice de co-occurence des niveaux de gris	Riaz et al. ; Redi et al.
Illumination		
Moyennes de Canaux de Luminance	Valeurs moyennes des canaux L^* ($L^*a^*b^*$) ou V (HSV)	Tong et al. ; Ke et al. ; Luo et Tang ; Khan et Vogel ; Males et al. ; Redi et al.
Asymétrie de la Luminance	Asymétrie des canaux L^* ($L^*a^*b^*$) ou V (HSV)	Fedorovskaya ; Li et Chen ; Redi et al.
Saturation des Pixels	Nombre de pixels dont la valeur de luminance est maximale	Males et al.
Modèles d'Illumination	Comparaison entre la luminance d'une image et la luminance d'images de bonne qualité esthétique	Khan et Vogel ; Redi et al.
Contraste		
Écart-Type de Canaux de Luminance	Écarts-Types des canaux L^* ($L^*a^*b^*$) ou V (HSV)	Khan et Vogel ; Xue et al. ; Redi et al.
Kurtosis de Canaux de Luminance	Kurtosis des canaux L^* ($L^*a^*b^*$) ou V (HSV)	Khan et Vogel
Longueur de l'histogramme	Longueur de 98% de la masse de l'histogramme	Ke et al. ; Wong et Low ; Aydin et al.
Formule de Michelson	$(L_{max} - L_{min})/(L_{max} + L_{min})$	Fedorovskaya ; Desnoyer et Wettergreen ; Khan et Vogel ; Redi et al.
Luminance RMS	$\sqrt{\frac{1}{MN}\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(L_{ij} - \bar{L})}$	Desnoyer et Wettergreen ; Males et al.
Réflectance	Somme des gradients du canal L^*	Fedorovskaya [Jiang et al. 2010]
Entropie	Entropie de l'image en niveaux de gris	Ng et al.
Égalisation	Distance entre l'image originale et l'image dont l'histogramme est égalisé	Redi et al.

Caractéristique	Méthode de calcul / Implémentation	Utilisé par...
Couleurs		
Histogrammes	Calcul d'histogrammes à partir des canaux de couleurs	Tong et al. ; Datta et al. ; Nishiyama et al. ; Desnoyer et Wettergreen ; Dhar et al. ; Kim et Kim ; Redi et al.
Nombre de Teintes	Nombre de couleurs distinctes dans les canaux de couleurs	Tong et al. ; Ke et al. ; Luo et Tang ; Pogačnik et al. ; Males et al. ; Xue et al.
Moyennes, Écarts Types	Moyennes et écarts types des différents canaux de couleur (H, a^*, b^*, etc.)	Tong et al. ; Ke et al. ; Datta et al. ; Riaz et al. ; Pogačnik et al. ; Males et al. ; Xue et al. ; Kim et Kim ; Redi et al.
Canal Sombre	Moyenne et écart-type du canal sombre : $I_{dark} = \min_{C \in R,G,B}(\min_{i' \in \Omega(i)}(I_C(i')))$	Tang et al.
Intensité des couleurs	Formule proposée par Hasler et Suesstrunk : Colorfulness $= \sqrt{Av(a^*)^2 + Av(b^*)^2} + 0{,}37\sqrt{StD(a^*)^2 + StD(b^*)^2}$	Aydin et al.

1.4.1.2 Mesures de niveau intermédiaire

Les mesures définies dans le tableau 1.5 tiennent compte de la composition de l'image. Elles sont souvent le résultat de l'exploitation d'informations connues sur les images (position du sujet, photo contenant un visage) ou d'informations extraites à partir de techniques de segmentation (avant, arrière-plan). Nous distinguons ici les mesures de segmentation de l'image, permettant d'obtenir des relations entre différentes régions de l'image (différences entre le sujet et l'arrière-plan), et les méthodes permettant d'obtenir directement des informations sur la composition d'une image (symétries, position de la ligne d'horizon, respect de la règle des tiers, etc.).

Tableau 1.5 – Caractéristiques de niveau intermédiaire utilisées dans l'état de l'art.

Caractéristique	Méthode de calcul / Implémentation	Utilisé par...
Extraction du sujet / Découpage de la photo		

Caractéristique	Méthode de calcul / Implémentation	Utilisé par...
Cartes de Saillance	Nombreux modèles proposés, basés sur l'estimation locale du contraste, de la netteté et de la coloration : [Itti et al. 1998 ; Siagian et Itti 2007]	Fedorovskaya ; Tong et al. ; Sun et al. ; Wong et Low ; Ng et al. ; Dhar et al. ; Xue et al. ; Kim et Kim
Estimation du Flou	Les zones floues correspondent à l'arrière-plan, les autres au sujet de l'image.	Loui et al. ; Luo et Tang ; Pogačnik et al. ; Tang et al. ; Aydin et al. ; Kim et Kim
Découpage par Blocs	Images découpées en grilles (souvent 3×3 ou 4×4) et des mesures sont calculées dans chaque bloc.	Datta et al. ; Desnoyer et Wettergreen ; Jiang et al. ; Riaz et al.
Redimensionnement	Images étudiées à différentes échelles, mesures calculées à chaque niveau.	Dhar et al. ; Marchesotti et Perronnin
Segmentation	Images segmentées automatiquement, mesures calculées dans chaque région.	Datta et al. ; Desnoyer et Wettergreen ; Bhattacharya et al. ; Tang et al.
Détection d'Objet	Extraction d'informations sur la position et la taille du visage dans l'image.	Li et al. ; Jiang et al. ; Khan et Vogel ; Pogačnik et al. ; Ravì et Battiato ; Males et al. ; Redi et al.
Composition d'une image		
Règle des Tiers	Positionnement du sujet principal par rapport aux points définis par la règle des tiers.	Datta et al. ; Luo et Tang ; Wong et Low ; Ng et al. ; Dhar et al. ; Khan et Vogel ; Pogačnik et al. ; Ravì et Battiato ; Males et al. ; Xue et al.
Profondeur de Champ	Différence de netteté entre le sujet et l'arrière-plan.	Datta et al. ; Luo et Tang ; Wong et Low ; Li et al. ; Dhar et al. ; Tang et al. ; Pogačnik et al. ; Males et al. ; Aydin et al.
Taille de l'image	Longueur + Largeur ou Longueur / Largeur	Datta et al. ; Khan et Vogel ; Riaz et al.
Taille des Régions	Proportion de l'image correspondant au sujet (avant-plan ou visage).	Li et al. ; Khan et Vogel ; Males et al.
Symétrie	Différences d'illumination ou de textures entre les parties gauche et droite de l'image ou du visage.	Ng et al. ; Jiang et al. ; Khan et Vogel ; Redi et al.

Caractéristique	Méthode de calcul / Implémentation	Utilisé par...
Lignes Directrices	Positions relatives des droites détectées par transformée de Hough.	Ng et al. ; Tang et al. ; Khan et Vogel
Horizon	Position et pente de l'horizon à l'aide de la transformée de Hough.	Ng et al. ; Tang et al. ; Khan et Vogel
Points de Fuite	Histogramme et positions des points de fuite.	Ng et al. ; Jiang et al.
Nombre de Régions	Complexité de l'image évaluée par le nombre de régions obtenues par segmentation.	Datta et al. ; Redi et al.
Points de repère dans le visage		
Positions	Abscisse et ordonnée de points décrivant les contours du visage et de ses attributs (sourcils, bouche, nez, yeux).	Todorov et Oosterhof ; Rojas et al.
Distances	Distances entre les positions de points de repère dans le visage.	Rojas et al.
Angles	Angles 3 points de repère adjacents.	Rojas et al.

1.4.1.3 Mesures de haut niveau

Les mesures décrites ici nécessitent un niveau d'abstraction supplémentaire par rapport à une étude directe sur les pixels de l'image, ce sont des notions qui se rapprochent plus du jugement humain. Nous distinguons les attributs qui sont associés à la photo (portrait, paysage, intérieur, ciel bleu...) et les attributs propres aux visages : le sexe, l'âge, la couleur des yeux, des cheveux, de la peau, la présence de lunettes, de barbe, etc. Le tableau 1.6 résume les différentes caractéristiques de haut niveau utilisées dans l'état de l'art.

Tableau 1.6 – Caractéristiques de haut niveau utilisées dans l'état de l'art.

Caractéristique	Méthode de calcul / Implémentation	Utilisé par...
Informations sur la photo		
Type de photographie	Portrait, photo d'animaux, de paysage, milieu urbain, etc.	Ng et al. ; Jiang et al. ; Dhar et al. ; Tang et al. ; Kim et Kim
Environnement de la photo	Intérieur, extérieur, présence de nuage, couleur du ciel, etc.	Dhar et al.

Caractéristique	Méthode de calcul / Implémentation	Utilisé par...
Harmonie des Couleurs	Comparaison entre les histogrammes de teinte de la photo considérée et de photos harmonieuses (modèle défini dans [Cohen-Or et al. 2006]).	Ke et al. ; Datta et al. ; Luo et Tang ; Ng et al. ; Desnoyer et Wettergreen ; Dhar et al. ; Pogačnik et al. ; Tang et al.
Familiarité	Comparaison entre une distribution de contours ou de couleurs avec des images / modèles de référence.	Ke et al. ; Datta et al. ; Khan et Vogel
Unicité	Comparaison entre le spectre de l'image et le spectre moyen d'une base d'images.	Redi et al.
Convexité	Rapport entre la taille du sujet et la taille de son enveloppe convexe.	Datta et al.
Informations sur le visage		
Orientation du visage	Angles de rotation possibles, placement du visage dans la photo...	Jiang et al. ; Li et al. ; Khan et Vogel ; Xue et al. ; Redi et al.
Ouverture des yeux	Évaluation de l'ouverture des yeux	Li et al.
Sourire	Détection de sourire	Li et al. ; Xue et al. ; Redi et al.
Expressions du visage	Apprentissage de modèles d'expressions faciales	Li et al. ; Ravì et Battiato
Relations entre visages	Cohérence entre différents visages d'une photo	Li et al. ; Xue et al.
Autres attributs	Modèles de couleurs de peau, de cheveux, des yeux, présence de lunettes, de barbe, de maquillage...	Redi et al.

1.4.2 Algorithmes d'apprentissage automatique

De nombreuses méthodes statistiques permettent de faire de la prédiction de catégories ou de la régression. Comme cela a déjà été évoqué, différentes méthodes produisent des résultats légèrement différents en termes de performances (meilleur taux de bonne classification, scores plus proches de la vérité terrain). Toutefois celles-ci sont très fortement conditionnées par les bases de photos considérées et par la fiabilité des annotations définissant la vérité terrain.

Dans cette partie, nous ne citons que des méthodes d'apprentissage automatique supervisé. En effet, celles-ci sont quasiment toujours employées dans l'estimation de la qualité esthétique de photographies. Nous ne rentrons pas dans les détails liés à l'implémentation de chaque algorithme. Toutes les méthodes présentées nécessitent au préalable un ensemble de photos

le plus vaste et varié possible, et chaque photo doit être associée à une catégorie, un score ou un ensemble de scores. L'apprentissage est alors guidé par les annotations, d'où la notion d'apprentissage supervisé. Les algorithmes ont pour tâche de concevoir automatiquement des modèles de prédiction de ces annotations à partir des caractéristiques mesurées sur chaque photo. Une fois l'apprentissage terminé, il est possible de prédire un score ou une catégorie correspondant à la qualité esthétique de photos ne figurant pas dans la base d'apprentissage.

Pour chaque méthode présentée, une brève description est donnée, et quelques paramètres importants sont abordés. Une liste non exhaustive de méthodes utilisées en estimation de qualité esthétique est donnée dans le tableau 1.7.

Tableau 1.7 – Méthodes d'apprentissage utilisées dans l'état de l'art.

Algorithme (Nom)	Principe et Paramètres	Utilisé par...
Régression Linéaire Multiple (MLR)	Estimation des paramètres liant les caractéristiques aux scores des images. Le modèle créé est dit linéaire car la fonction reliant les variables aux scores réels est linéaire. Les méthodes d'estimation des paramètres peuvent varier (moindres carrés, maximum de vraisemblance).	Fedorovskaya ; Tong et al. ; Datta et al. ; Khan et Vogel
Classification Bayésienne (BC)	Classification basée sur le théorème de Bayes. Une BC naïve suppose une forte indépendance entre les caractéristiques, représentées chacune par une distribution gaussienne.	Tong et al. ; Ke et al. ; Datta et al. ; Luo et Tang
Méthode des K plus proches voisins (KNN)	Les K photos d'apprentissage dont les caractéristiques sont les plus semblables à celles de la photo à tester sont prises en compte pour la classification. Prédire un score est possible en moyennant les scores fournis par les plus proches voisins. Le calcul de distance, le nombre de voisins et leur influence respective peuvent être ajustés.	Ke et al. ; Khan et Vogel
Boosting	Des classifieurs faibles (par exemple une seule caractéristique) sont combinés afin de construire un classifieur plus robuste. La méthode de combinaison et de sélection de poids peut fortement varier et de nombreux algorithmes de boosting existent : AdaBoost et ses variantes, RankBoost, etc. Le choix du nombre et de la nature des classifieurs faibles utilisés est déterminant.	Tong et al. ; Luo et Tang ; Desnoyer et Wettergreen ; Khan et Vogel ; Faria et al.

Algorithme (Nom)	Principe et Paramètres	Utilisé par...
Machines à Vecteurs de Support (SVM)	Classifieur dont le but initial est de séparer deux groupes de données dans l'espace des caractéristiques. Des garanties sont données pour maximiser la distance entre les données et la frontière de séparation (marge maximale, définie par la position des vecteurs de support). Il est possible de traiter les cas de données non linéairement séparables en projetant les données dans des espaces de plus grande dimension à l'aide d'une fonction noyau. Un noyau gaussien produit généralement de bons résultats. La méthode peut être étendue à une classification à plus de deux catégories (en combinant différentes séparations), ainsi qu'à la régression.	Tong et al. ; Datta et al. ; Luo et Tang ; Wong et Low ; Ng et al. ; Li et al. ; Jiang et al. ; Dhar et al. ; Marchesotti et Perronnin ; Khan et Vogel ; Pogačnik et al. ; Tang et al. ; Males et al. ; Xue et al. ; Redi et al.
Réseau de Neurones Artificiels (ANN)	Automate transformant un ensemble de caractéristiques d'entrées en sortie précise. Les réseaux couramment utilisés en classification et régression propagent les informations uniquement vers l'avant du réseau (pas de boucle). Ils sont souvent constitués de 3 couches : une d'entrée (les caractéristiques), une de sortie (la ou les classes de l'image), et une couche cachée qui sert à combiner les entrées. Le nombre d'unités du réseau (les neurones), les fonctions d'activation de ces neurones ainsi que les algorithmes d'apprentissage peuvent varier.	Wong et Low ; Riaz et al.
Forêt d'arbres de décision (RF)	Ensemble d'arbres décisionnels construits à partir d'un échantillonnage aléatoire des caractéristiques et des images de la base d'apprentissage. Chaque arbre a une influence sur le résultat de la prédiction : un vote (classification) ou un score (régression). Le nombre d'arbres, leur forme (profondeur, caractéristiques...) sont des paramètres à définir.	Loui et al. ; Khan et Vogel ; Faria et al. ; Kim et Kim

Chaque algorithme possède des avantages et inconvénients, selon le type de données à exploiter ou le temps de calcul disponible. Il ressort toutefois qu'un très grand nombre de travaux utilisent les machines à vecteurs de support. Plusieurs raisons expliquent cela. Tout d'abord, la méthode est plutôt facile d'utilisation : de très nombreuses bibliothèques d'apprentissage automatique implémentent cet algorithme. L'une d'entre elles est décrite en détails par [Chang et Lin 2011]. Peu de paramètres sont à définir et ceux présents par défaut sont généralement suffisants pour l'obtention de bons résultats. Les SVM permettent d'effectuer aussi bien de la

classification que de la régression, et les résultats obtenus sont en général aussi bons que ceux obtenus par d'autres algorithmes.

Plusieurs autres méthodes sont de plus en plus largement utilisées dans les travaux récents. Les forêts d'arbre de décision, introduites par les travaux décrits par [Breiman 2001], sont actuellement considérées pour l'évaluation de la qualité esthétique pour au moins 3 raisons. En effet, l'idée de prendre une décision reposant sur les résultats de plusieurs agents (les arbres de décisions) introduit une certaine robustesse aux données très bruitées. Ensuite, la possibilité de paralléliser les instructions à l'aide de plusieurs processeurs concurrents, couplée à l'indépendance de chaque arbre de décision, fait que l'apprentissage et la prédiction à l'aide d'arbres de décisions est très rapide. Enfin, les forêts d'arbre de décision, de par leur prédiction finale consensuelle, tendent à reproduire les jugements humains. Ainsi, chaque arbre possède des critères d'évaluation différents, et de la même façon que la vérité terrain est obtenue par la moyenne des différentes évaluations humaines, la prédiction de la forêt représente la moyenne des évaluations subjectives de chaque arbre.

Actuellement, les réseaux de neurones sont de plus en plus couramment utilisés dans une grande partie des tâches liées à l'apprentissage automatique. Toutefois, l'architecture des réseaux actuels devient très complexe (couches de convolutions en entrée, nombreuses couches cachées intermédiaires) et l'apprentissage requiert la disponibilité d'un grand nombre d'images annotées, ce qui est difficile à obtenir dans le cas où les annotations sont issues d'un grand nombre de jugements humains.

Finalement, il existe de très nombreux algorithmes d'apprentissage automatique adaptés à l'estimation de qualité esthétique de photographie. Si les performances sont très fortement dépendantes des caractéristiques et des bases de données, l'impact du choix de l'algorithme peut également être important.

1.5 Conclusion

De par son caractère subjectif, l'estimation automatique de la qualité esthétique de photographies est une tâche complexe. Il faut en effet :

1. avoir une idée des critères permettant l'évaluation des photos. Ces critères peuvent dépendre des individus, des conditions de visionnement, etc.
2. savoir estimer automatiquement ces critères. Cette tâche peut s'avérer très complexe lorsqu'il s'agit de critères abstraits : composition, sujet de la photo.
3. constituer une vérité terrain fiable. Cela passe par la création d'une base de données variée et conséquente, représentative de l'ensemble des photos, et par l'évaluation de cette base par des humains. Créer une telle base de données est un travail difficile et laborieux.
4. combiner les mesures effectuées sur les photos avec les résultats de la vérité terrain à l'aide d'algorithmes d'apprentissage automatique. L'obtention de résultats satisfaisants est dépendante du choix de l'algorithme et de ses paramètres.

Actuellement, un grand nombre de travaux ont été réalisés concernant le problème de la distinction entre des photos de très haute et de très basse qualité. Les résultats sont encourageants : selon les caractéristiques et les bases de données utilisées, les résultats varient entre 60 et 90% de bonne classification. Le problème de la régression est moins souvent abordé, car plus difficile. Les valeurs associées aux photos ne sont plus des étiquettes de classe, mais des nombres réels, dont les valeurs sont souvent toutes différentes. L'apprentissage a donc besoin d'un ensemble de photos plus conséquent. Une première méthode de régression est proposée par [Datta et al. 2006], mais ne s'applique pas au cas particulier des photos contenant des visages. [Li et al. 2010a] étudient les photos contenant des personnes et présentent une méthode de régression dont l'erreur quadratique moyenne entre la prédiction et la vérité terrain est réduite de 25% par rapport à celle d'une évaluation aléatoire. Ce résultat dépasse largement les prédictions réalisées par [Datta et al. 2006], ce qui prouve bien le caractère particulier des photos contenant des visages.

À notre connaissance, l'étude spécifique et automatique de la qualité esthétique de photos de visage (voir les contraintes définies sur la figure 3 de l'introduction) n'est abordée qu'à partir des travaux de [Males et al. 2013 ; Redi et al. 2015]. Dans la suite de ce document, nous focalisons notre étude sur ces photos particulières. De nouvelles bases de photos sont présentées, des caractéristiques adaptées aux photos de visage sont calculées et des méthodes d'évaluation sont proposées, tant pour le problème de la classification que pour la régression.

Les difficultés rencontrées pour l'évaluation des impressions de compétence et de sympathie sont les mêmes que celles évoquées pour l'étude de la qualité esthétique des photos de visage : évaluations subjectives, manque de données. En outre, les caractéristiques à prendre en compte sont complexes à appréhender car elles font appel à des notions subjectives, liées aux émotions et à l'interprétation de chacun. Dans ce document, des pistes sont proposées afin d'ajouter aux estimations de qualité esthétique des indices permettant d'évaluer les impressions de compétence et de sympathie. Il ne s'agit donc pas de proposer un modèle complet intégrant tous les paramètres mais d'obtenir un résultat plus précis que les modèles de l'état de l'art et qui s'approche suffisamment des évaluations humaines sur des visages réels.

La figure 1.1 résume les différentes informations que nous extrayons sur les photos. Nous rappelons sur cette figure que nous travaillons sur des photographies de visage, et que nous extrayons différents types d'informations :

— Des informations permettant de quantifier différents aspects de la qualité esthétique des photographies (mesures de texture, de couleur, de contraste, etc.).
— Des informations sur le sujet de l'image. Dans nos travaux, nous tenons compte des relations entre les différentes régions de l'image (avant-plan et arrière-plan) en calculant des caractéristiques décrivant différents aspects de la qualité esthétique des photographies (mesures de texture, de couleur, de contraste, etc.) dans différentes régions de la photo (le visage, les yeux, la bouche).
— Des informations sur la composition de l'image. Comme nous travaillons uniquement sur des photographies de visage, les seules informations relatives à la composition que nous considérons sont la position et la taille du visage. Nous n'intégrons pas d'autres informations sur la composition (par exemple la présence d'une ligne d'horizon) des

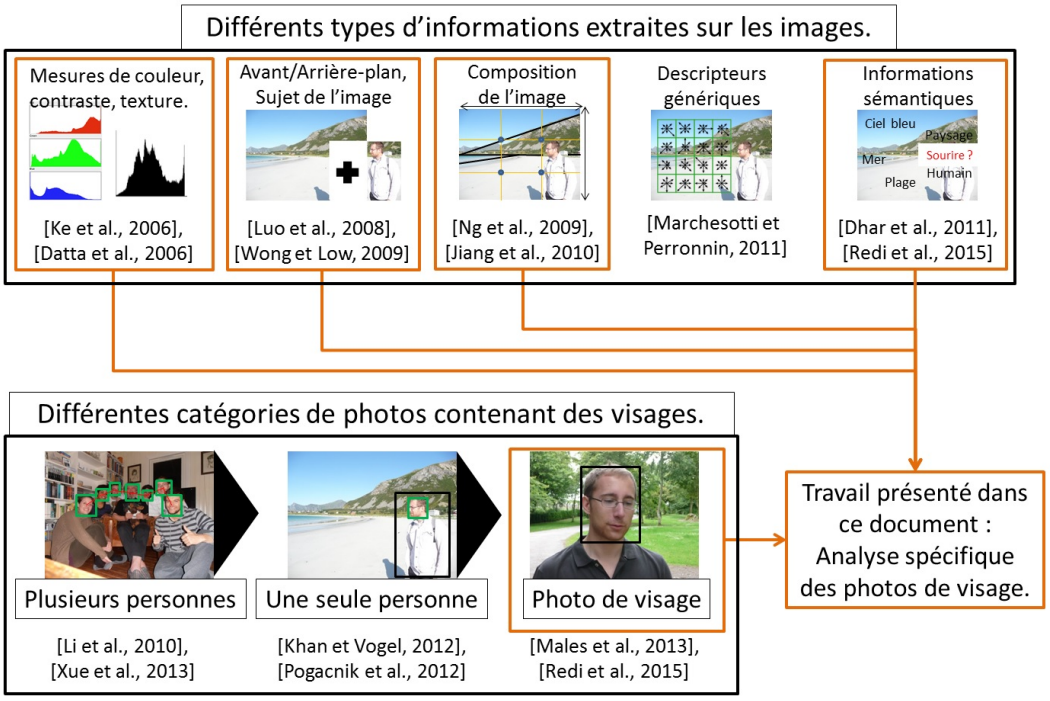


FIGURE 1.1 – Bilan des caractéristiques et des différentes catégories d'images évoquées. Celles considérées dans ce travail sont encadrées.

photos.

— Des informations sémantiques sur le visage. Ces informations sont utilisées essentiellement pour l'évaluation des impressions de compétence et de sympathie.

Chapitre 2

Analyse des données et apprentissage supervisé

Sommaire

2.1 Introduction 43
2.2 Prétraitements - Gestion des données 46
2.2.1 Représentation des caractéristiques 46
2.2.2 Normalisation des données 47
2.2.3 Concaténation des données 48
2.3 Sélection des données pertinentes - L'algorithme Relief 49
2.3.1 Principe 49
2.3.2 Algorithme Relief - Classification binaire 50
2.3.3 Algorithme ReliefF - Classification généralisée 54
2.3.4 Algorithme RReliefF - Régression 56
2.4 Apprentissage automatique et prédiction 57
2.4.1 Algorithmes d'apprentissage 57
2.4.2 Protocole et critères d'évaluation 67
2.5 Post-traitement - Fusion des scores 71
2.5.1 Normalisation des scores pour la régression 71
2.5.2 Fusion des scores pour la classification et la régression 72
2.6 Conclusion 75

2.1 Introduction

Dans ce chapitre, nous définissons les principales étapes qui permettent la création des modèles établis dans ce document. Nous proposons ainsi un cadre de travail générique, très largement inspiré des précédents travaux sur l'estimation de la qualité esthétique de photographies. Ce cadre est adapté et amélioré afin de tenir compte des problématiques liées aux photos de visage (peu d'images annotées disponibles, évaluations très subjectives). Il est repris dans les chapitres suivants, où les méthodes proposées sont mises en œuvre et évaluées.

L'approche proposée est dite descendante (ou "top-down") car nous partons de l'objectif, définissons les attributs à évaluer en fonction de nos connaissances en photographie et de notre

intuition, puis cherchons à construire un modèle à partir des caractéristiques extraites. Nous avons choisi cette approche pour plusieurs raisons. Tout d'abord, elle correspond à la façon dont raisonnent les humains : est-ce que cette image est floue ? Si non, est-elle bien éclairée ? Les couleurs sont-elles bien choisies ? C'est en prenant en compte ces différents critères et en les combinant que nous prenons nos décisions. Une image floue et bien contrastée est-elle plus jolie qu'une image nette et mal contrastée ? En extrayant des caractéristiques permettant de répondre à chacune de ces questions, puis en les combinant à l'aide d'algorithmes d'apprentissage, nous créons un modèle permettant d'estimer la qualité esthétique de photographies. Cette approche est également intéressante car elle rend accessible les informations liées à la prise de décision : si l'image est mal évaluée, il suffit de regarder les valeurs de chaque caractéristique et d'en déduire celles dont les valeurs correspondent à une image faiblement évaluée. Enfin, la grande majorité des travaux sur la qualité esthétique de photos repose sur une approche semblable et propose des modèles pertinents, qui devraient pouvoir s'appliquer au cas particulier des photos de visage.

A l'inverse, une approche ascendante correspond par exemple à l'utilisation par les algorithmes d'apprentissage de descripteurs d'image génériques (par exemple SIFT, SURF). Si les performances de telles méthodes sont comparables aux approches descendantes (voir [Marchesotti et Perronnin 2011]), et que la plupart des informations (netteté, contraste, etc.) peuvent être implicitement encodées dans ces descripteurs, ceux-ci sont surtout utilisés lorsque les images considérées sont très variées : portraits, paysages, animaux, architecture, etc. Dans le cas particulier des photos de visage, les attributs liés à la qualité esthétique de la photo (distinction entre arrière-plan et visage, position du visage, éléments du visage) sont très précis et donc difficiles à encoder directement et efficacement à l'aide de descripteurs génériques. Une autre approche ascendante possible est l'utilisation de Réseaux de Neurones Convolutionnels [LeCun et al. 1998] (*CNN*), afin d'apprendre automatiquement les filtres de convolution pertinents pour l'estimation de qualité esthétique. Toutefois, si ces réseaux se sont montrés très efficaces en reconnaissance d'objet ou classification d'images, leur utilisation requiert de très nombreuses données pour créer un modèle de prédiction robuste. Or dans notre cas, nous n'avons à notre disposition que quelques centaines ou milliers d'images disponibles, ce qui est bien inférieur aux données généralement utilisées pour ce type d'algorithme.

Nous avons également fait le choix d'utiliser un algorithme d'apprentissage au lieu de définir directement une formule explicite à l'aide de notre intuition et des données disponibles, contrairement aux travaux de [Battiato et al. 2013 ; Aydin et al. 2015]. En effet, si une formule explicite combinant différents attributs peut constituer une première approximation de la qualité esthétique, il est difficile d'obtenir une évaluation précise à cause du caractère subjectif de cet objectif. Nous supposons donc qu'il n'existe pas de formule explicite permettant de résoudre notre problème, et tentons de créer un modèle s'approchant de la vérité terrain à partir d'un apprentissage statistique.

Dans ce chapitre, nous ne cherchons pas à savoir quelles sont les caractéristiques discriminantes d'une photo réussie. Nous détaillons les étapes d'extraction des caractéristiques dans les chapitres 3 et 4, respectivement dédiés à l'évaluation de la qualité esthétique et à l'estimation des impressions de compétence et de sympathie suggérées par les photos de visage.

Nous présentons dans ce chapitre un prétraitement possible des données (qui sont les caractéristiques extraites sur les photos) en 2.2. Cette analyse correspond à la normalisation de ces données et à la façon dont celles-ci sont combinées avant la phase d'apprentissage (concaténation dans un unique vecteur). Une éventuelle réduction de la dimension du vecteur de caractéristiques est détaillée en 2.3. Celle-ci constitue une des améliorations que nous proposons par rapport aux méthodes classiques. Une fois toutes les données extraites et organisées, nous abordons la phase d'apprentissage en section 2.4, où sont détaillés les 4 algorithmes d'apprentissage considérés et testés ainsi que les différents critères d'évaluation des performances. Enfin, nous exposons en section 2.5 une nouvelle méthode permettant d'augmenter la robustesse et la précision de la prédiction, en fusionnant les sorties de différents algorithmes. Le plan de ce chapitre est résumé sur la figure 2.1.

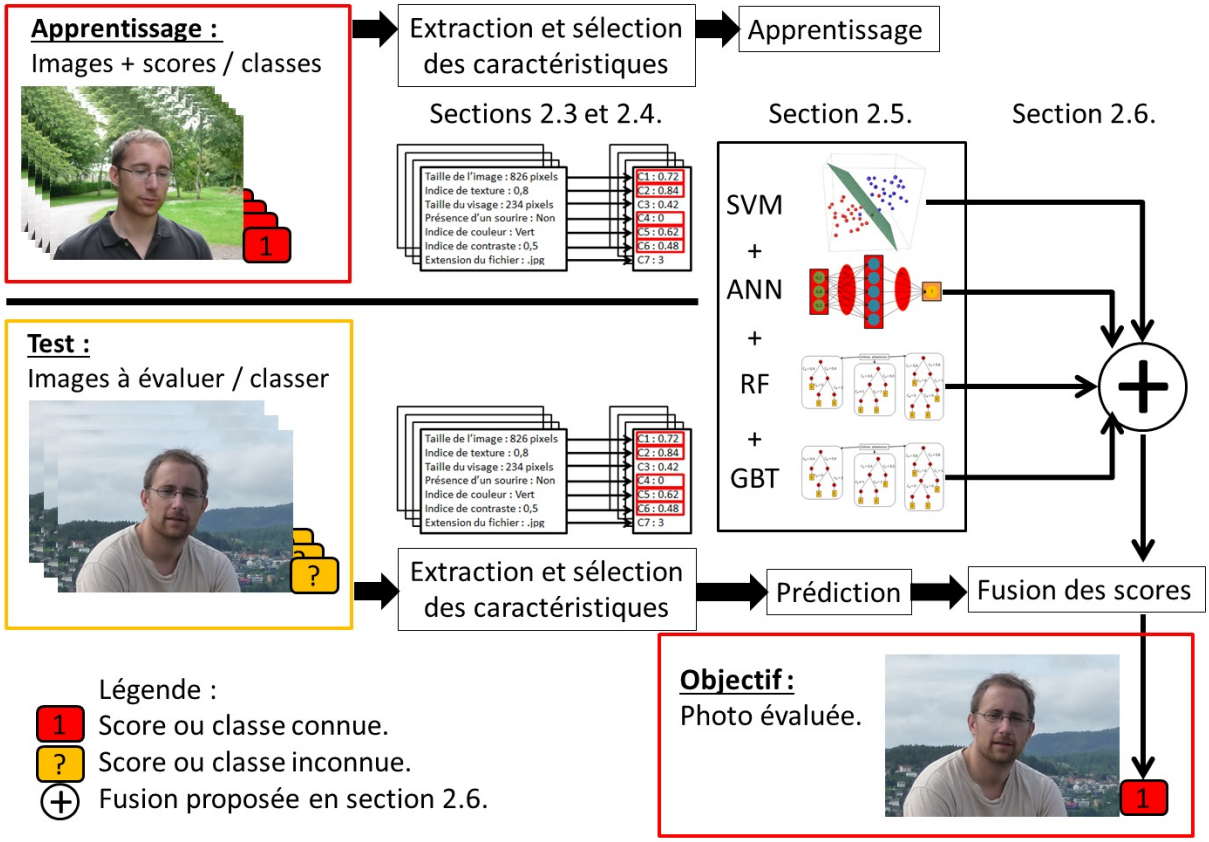


FIGURE 2.1 – Schéma présentant les différentes étapes permettant l'évaluation des images. Un premier ensemble d'images (partie haute de la figure) permet d'apprendre différents modèles, le second (partie basse de la figure) est utilisé pour tester ces modèles.

Nous pouvons finalement résumer les différentes contributions apportées par nos travaux et présentées dans ce chapitre par les 2 points suivants :

— Utilisation d'un algorithme permettant une sélection pertinente des caractéristiques, qui augmente la précision de la prédiction tout en réduisant le temps de calcul (section 2.3).

— Fusion des sorties proposées par les différents algorithmes d'apprentissage, permettant

une augmentation de la robustesse et de la précision des prédictions (section 2.5.2).

2.2 Prétraitements - Gestion des données

Après avoir donné un éventail assez large des différents descripteurs considérés pour l'évaluation des photos de visage (cf. le chapitre 2), il nous faut définir une façon d'extraire, de quantifier et de combiner les différentes valeurs encodant ces descripteurs. Nous distinguons essentiellement 3 étapes décrites sur la figure 2.2 : représentation des caractéristiques, normalisation, puis sélection des caractéristiques les plus pertinentes. Cette dernière étape est détaillée dans la section 2.3.

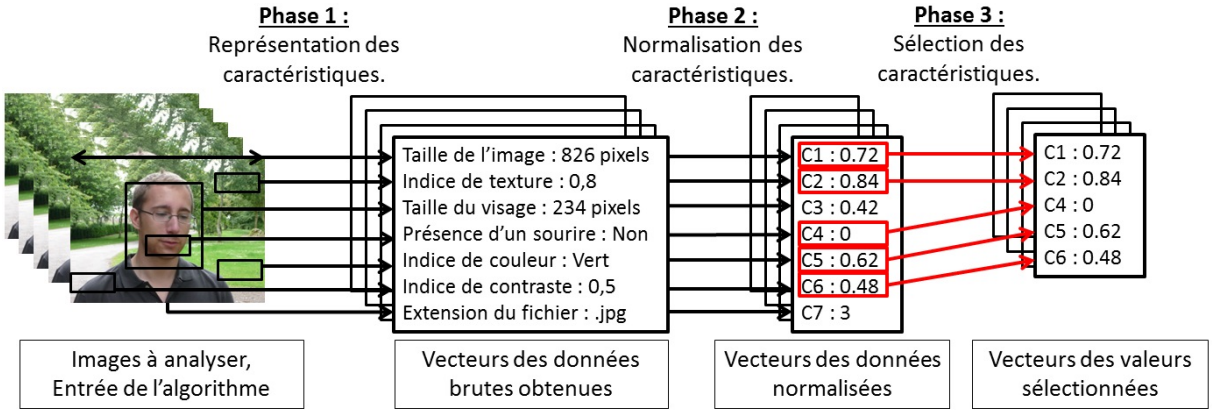


FIGURE 2.2 – Principales étapes liées à l'analyse des images. Les détails concernant l'extraction des caractéristiques sont donnés dans les chapitres suivants ; la méthode de sélection de caractéristiques est expliquée en section 2.3.

Dans ce chapitre, nous ne détaillons pas la phase d'extraction des caractéristiques, car celle-ci est dépendante de l'application visée, alors que nous souhaitons conserver la généricité de la méthode (quelques exemples sont donnés sur la figure 2.2). Nous traitons dans cette section de la représentation des données, de leur normalisation, puis nous discutons de la façon dont elles sont fusionnées.

2.2.1 Représentation des caractéristiques

Nous distinguons essentiellement les descripteurs à valeurs continues et à valeurs discrètes. Les descripteurs à valeurs continues peuvent être représentés par des nombres réels, dont l'ensemble des valeurs possibles est défini par un intervalle (entre 0 et 1, entre -1 et 1, de 0 à l'infini, etc.). Ces données correspondent généralement à des mesures globales et génériques : calcul du contraste, d'indices de netteté tels que la moyenne des gradients, etc. Les descripteurs à valeurs discrètes peuvent être encodés à l'aide d'étiquettes représentées par des nombres entiers, dont l'ensemble des valeurs possibles est fini. Ces valeurs traduisent souvent des notions

plus complexes à calculer directement à partir des pixels. Est-ce un homme ou une femme ? La personne sourit-elle ? Porte-t-elle des lunettes ? Si oui, des lunettes de soleil ? Ces informations peuvent éventuellement être complétées par des données continues, correspondant à une probabilité de certitude ou à un niveau d'expression. Par exemple, un algorithme capable de détecter un sourire peut également donner un indice d'intensité de ce sourire. La différence entre lunettes et lunettes de soleil étant parfois difficile, les algorithmes peuvent aussi renvoyer un indice de fiabilité : "La probabilité que ces lunettes soient des lunettes de soleil est de 80%".

La distinction entre ces deux types de caractéristiques est cruciale car celles-ci ne peuvent pas forcément être traitées de la même manière par les algorithmes d'apprentissage. Ainsi, il est important de normaliser les données continues pour réduire les effets d'échelle lors de l'apprentissage : les mesures dont les valeurs varient entre 0 et 1 risquent d'avoir très peu de poids face à celles dont les valeurs ne sont pas bornées. Notons également que les algorithmes d'apprentissage ne se comportent pas tous de la même façon face à des données discrètes. Enfin, les méthodes de sélection de caractéristiques ou de réduction de dimension sont également sensibles à ces différences.

Un autre aspect important concernant la représentation des caractéristiques est la gestion des données manquantes. En effet, s'il est toujours possible d'évaluer le contraste ou la netteté, il est par exemple impossible de quantifier la probabilité que des lunettes soient des lunettes de soleil si la personne ne porte pas de lunettes. Or lors de l'apprentissage, les vecteurs représentant les images doivent tous avoir la même taille (les informations sont stockées sous la forme d'une matrice dont les lignes correspondent aux caractéristiques d'une photo), il n'est pas possible de simplement omettre une donnée. Plusieurs stratégies de "remplissage" sont alors possibles. Une première idée consiste à attribuer une valeur moyenne (une probabilité de 0,5 pour l'exemple des lunettes) ou aléatoire (selon une loi à définir) aux données manquantes. Une seconde est l'attribution d'une valeur arbitraire signalant l'absence de la donnée. Nous adoptons la seconde solution car certains des algorithmes que nous utilisons sont capables de tirer profit des données manquantes : l'impossibilité de mesurer l'information est une donnée à part entière.

2.2.2 Normalisation des données

Les données représentées par des valeurs réelles (variables continues) sont généralement normalisées, de façon à ce que chaque caractéristique ait à peu près le même ordre de grandeur. Par exemple, le nombre de pixels d'une image, qui peut donner un indice sur la qualité de l'image, peut être de l'ordre de plusieurs millions. A l'inverse, certaines mesures, telles que le contraste, prennent des valeurs plus petites, comprises entre 0 et 1. Il est donc nécessaire de ramener les deux caractéristiques à des valeurs comparables pour que celles-ci aient le même poids par la suite. Ceci est particulièrement important lors de l'utilisation d'algorithmes tels que les réseaux de neurones. En effet, lors de l'apprentissage, les paramètres du réseau associés aux variables aux plus grandes valeurs sont plus rapidement modifiés car de petites différences relatives introduisent d'importantes erreurs de prédiction. La convergence de l'algorithme est alors ralentie et a plus de chances de s'achever sur un minimum local. Pour remédier à ce

problème, dans toutes nos expériences, nous appliquons le traitement suivant à nos variables avant la phase d'apprentissage.

Soit f une variable représentée par des valeurs réelles quelconques. Soit $E = (f_1, \ldots, f_N)$ l'ensemble des valeurs de f mesurées sur les N images de la base d'apprentissage, et notons f_{min} et f_{max} les valeurs minimale et maximale de E. L'ensemble E est modifié à l'aide de la formule suivante, de façon à ce que chaque valeur de f soit comprise entre 0 et 1 :

$$E_{\text{Corrigé}} = (\frac{f_1 - f_{min}}{f_{max} - f_{min}}, \ldots, \frac{f_N - f_{min}}{f_{max} - f_{min}}) \tag{2.1}$$

Les valeurs de chaque variable f sont ainsi toutes incluses entre 0 et 1 et réparties à l'intérieur de cet intervalle de manière à ce qu'au moins une valeur soit égale à 0 et à 1 (respectivement f_{min} et f_{max}). Les valeurs f_{min} et f_{max} obtenues pour chaque variable sont conservées et utilisées afin de normaliser les données lors de la phase de prédiction.

De nombreuses autres transformations sur les caractéristiques sont possibles. Celles-ci peuvent par exemple être centrées ou réduites :

— Une variable est dite **centrée** lorsque l'on a retranché sa moyenne à chacune de ses valeurs. L'espérance d'une variable centrée est ainsi toujours fixée à 0. Nous n'avons pas constaté de changement significatif en centrant les variables dans nos expériences.

— Une variable est **centrée réduite** lorsqu'elle est centrée, et que chacune de ses valeurs a été divisée par l'écart-type de la variable. L'écart-type et la variance d'une variable centrée et réduite sont égaux à 1. Dans nos expériences, réduire les variables a tendance à amplifier l'impact de différences non significatives entre deux valeurs, car les données sont fortement bruitées et que l'écart-type des variables est généralement très faible.

Nous avons choisi de n'appliquer que la normalisation décrite par l'équation 2.1 pour nos données, de manière à ce que les valeurs soient comprises entre 0 et 1. En effet aucune amélioration de performance n'a été constatée lorsque les données sont centrées ou réduites. Cette étape a pour objectif de pouvoir comparer différentes variables dont les ordres de grandeur sont très différents.

Remarque : Nous n'avons traité que du cas des valeurs continues, car les caractéristiques discrètes ne sont pas modifiées. Celles-ci sont directement représentées par les entiers de 0 à V, où V est le nombre de valeurs distinctes que peut prendre la variable.

2.2.3 Concaténation des données

Nous considérons maintenant un ensemble de variables normalisées, décrivant chacune un aspect de l'image. Il existe de nombreuses méthodes permettant de fusionner ces données afin d'obtenir une évaluation globale de l'image. Nous en distinguons deux types : celles dont la combinaison des informations s'effectue avant l'apprentissage (fusion précoce), et celles dont la combinaison s'effectue après (fusion tardive).

Le principe de la fusion précoce est de concaténer toutes les valeurs des variables dans un unique vecteur représentant l'image. Ce vecteur est alors fourni en entrée de l'algorithme

d'apprentissage pour chaque image. Le modèle est construit en adaptant les paramètres associés à chaque élément du vecteur aux évaluations subjectives de chaque image. Cette méthode est particulièrement adaptée à notre problème car l'évaluation des photos prend en compte un grand nombre de critères, et il n'est pas possible de proposer un modèle pertinent d'évaluation de qualité esthétique ou du visage sans tenir compte de tous les éléments. La fusion précoce ne requiert qu'un seul apprentissage et les images n'ont besoin d'être annotées que selon un unique critère (visage sympathique, image de bonne qualité esthétique, etc.).

A l'inverse, la fusion tardive nécessite un apprentissage particulier pour chaque variable, ou au moins chaque type de variables, afin de créer un modèle représentant un aspect particulier de l'image : netteté, contraste, sourire, etc. Les modèles sont ensuite combinés pour former un modèle global. Cette méthode présente l'avantage de fournir des informations précises pour chaque critère : l'image contient de très belles couleurs, la personne sourit, mais la photo est floue. Étant donné que la construction de chaque modèle nécessite des annotations plus complexes sur les images, donnant des informations sur chaque descripteur, nous nous limitons à l'approche basée sur une fusion précoce dans nos expériences. Toutefois, un exemple d'utilisation de fusion tardive est proposé au chapitre 4, où le modèle d'estimation de qualité esthétique défini au chapitre 3 est combiné au modèle d'évaluation de la sympathie proposé au chapitre 4.

2.3 Sélection des données pertinentes - L'algorithme Relief

2.3.1 Principe

Il est possible d'extraire un très grand nombre d'informations à partir d'une image. Ces données, une fois normalisées et concaténées dans un unique vecteur décrivant l'image, représentent des informations très variées. Celles-ci peuvent être très pertinentes pour une certaine application, et complètement inutiles pour d'autres. Par exemple, une photo où les yeux sont fermés est généralement considérée comme ratée, mais ce n'est pas forcément un critère de mauvaise qualité esthétique. Certaines variables sont donc plus pertinentes que d'autres. Dans le cas d'une photo de visage, la netteté globale de la photo importe peu : seul le sujet de l'image doit être mis en valeur à l'aide d'un contraste élevé et de contours nets. Il est ainsi intéressant de définir des mesures permettant d'ordonner les variables selon leur importance, ce qui permet de :

— Voir quelles sont les données les plus pertinentes,
— Améliorer les méthodes d'extraction de caractéristiques,
— Supprimer les caractéristiques jugées non pertinentes.

En ne conservant que des caractéristiques pertinentes dans les modèles finaux, nous pouvons améliorer la précision des évaluations sur les images. Dans cette section, nous détaillons l'algorithme Relief, que nous utilisons afin d'obtenir des informations sur la pertinence de chaque caractéristique.

Différentes méthodes peuvent fournir des informations sur la pertinence de chaque carac-

téristique. Par exemple la fusion tardive, discutée en 2.2.3, permet d'évaluer les performances de chacune des caractéristiques séparément. Ainsi, les caractéristiques les plus pertinentes sont celles correspondant aux performances les plus élevées. La fusion tardive requiert cependant un grand nombre d'apprentissages spécifiques à chaque caractéristique, ce qui prend du temps. Nous avons finalement choisi l'algorithme Relief car il a été utilisé et testé dans de nombreux travaux antérieurs, et appliqué avec succès à l'estimation de la qualité esthétique de photographies [Pogačnik et al. 2012]. De plus, cet algorithme est capable de s'adapter aux différents problèmes que sont la classification et la régression, et permet de travailler simultanément avec des données continues et discrètes. L'objectif de l'algorithme Relief est de donner un indice de la capacité d'une caractéristique à distinguer des images aux caractéristiques semblables (proches dans l'espace des caractéristiques) mais dont l'évaluation est différente (classe ou score différents). L'algorithme ainsi que les différents paramètres que nous avons choisis sont expliqués dans les sections suivantes. Nous y présentons les différentes versions proposées par [Robnik-Šikonja et Kononenko 2003] et montrons comment adapter l'algorithme à nos données.

2.3.2 Algorithme Relief - Classification binaire

Principe

Nous nous intéressons dans un premier temps au problème de la classification binaire. Chaque image est associée à une des deux catégories, que nous notons c_1 et c_2. Dans le cas de l'estimation de la qualité esthétique, c_1 et c_2 sont respectivement les ensembles des images de bonne et de mauvaise qualité esthétique. L'objectif est de quantifier la capacité de chaque variable à fournir une indication permettant de retrouver la classe à laquelle appartient l'image.

Prenons une image i dans la catégorie c_1, et cherchons les deux images $i_1 \in c_1$ et $i_2 \in c_2$ dont les caractéristiques sont les plus semblables à celles de i. Soient respectivement f_i, f_{i_1} et f_{i_2} les valeurs de la caractéristique $\mathcal{F}$ pour les images i, i_1 et i_2. Si f_i et f_{i_1} sont proches, alors la caractéristique est probablement pertinente. A l'inverse, si f_i et f_{i_2} sont proches, alors la caractéristique est probablement peu pertinente. Cette idée de base permet de construire une première estimation de la pertinence de $\mathcal{F}$ (notée $\mathcal{R}(\mathcal{F})$) : $\mathcal{R}(\mathcal{F}) = |f_i - f_{i_2}| - |f_i - f_{i_1}|$. Dans le cas de variables continues réparties dans l'intervalle [0,1], une valeur élevée de $\mathcal{R}(f)$ (proche de 1) suggère que la caractéristique est très discriminante, tandis qu'une valeur faible (proche de 0 voire inférieure à 0) indique une caractéristique non discriminante. L'algorithme complet consiste à réitérer cette opération, résumée sur la figure 2.3 en sélectionnant des images i aléatoirement, en trouvant leurs plus proches voisins dans chaque classe puis en mettant à jour l'indice : $\mathcal{R}(\mathcal{F}) = \mathcal{R}(\mathcal{F}) + |f_i - f_{i_2}| - |f_i - f_{i_1}|$. Cependant, certains points de l'algorithme sont toujours à éclaircir : quelles mesures de distance utiliser entre les images ? Entre les caractéristiques ? Combien d'itérations réaliser ?

Distance entre images

Afin de trouver les images les plus semblables à une image de référence, il faut définir une

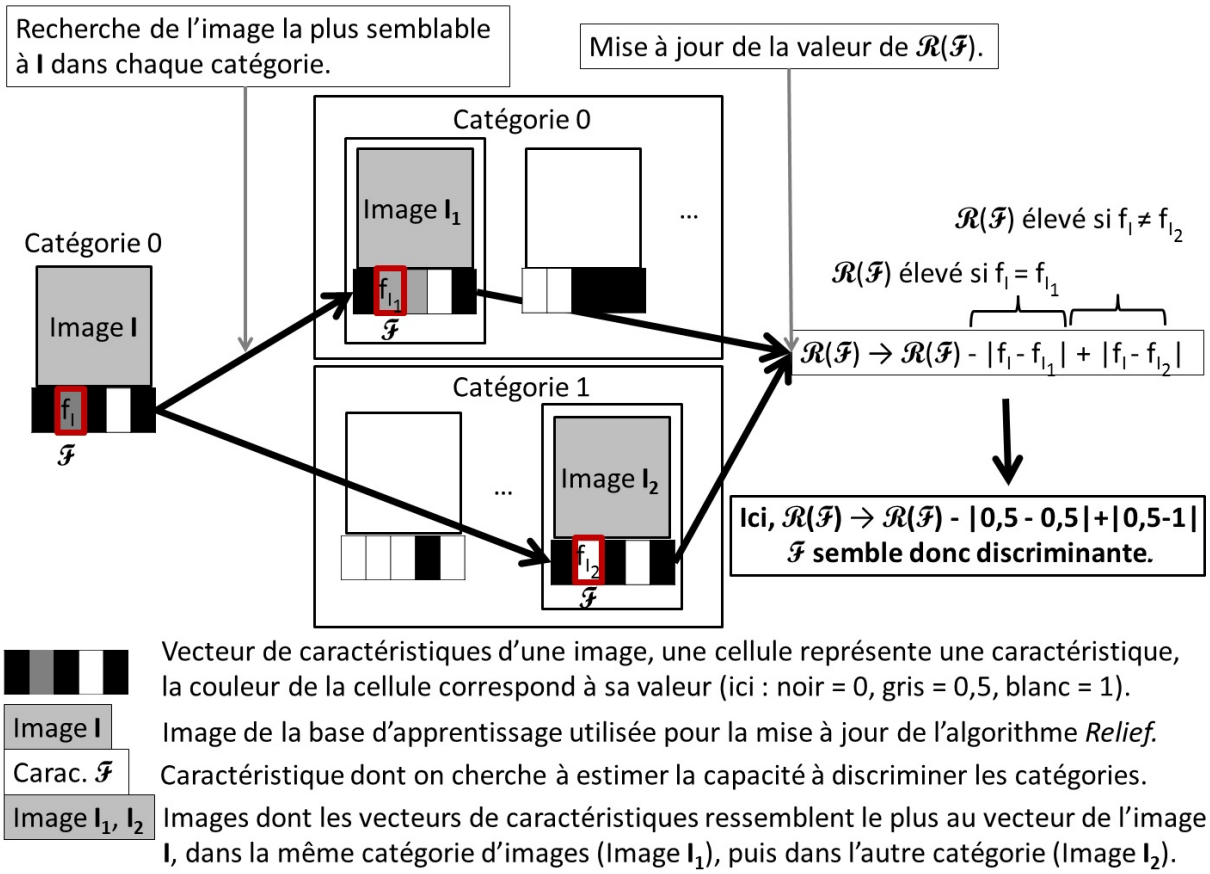


FIGURE 2.3 – Une itération de l'algorithme Relief. Cette opération est répétée de façon à obtenir une estimation fiable de la capacité de discrimination de la caractéristique $\mathcal{F}$.

mesure de distance entre deux images. Généralement, les distances entre images sont calculées dans l'espace des caractéristiques. Les métriques couramment utilisées sont la distance euclidienne et la distance de Manhattan ; [Robnik-Šikonja et Kononenko 2003] indiquent que les performances de l'algorithme sont similaires pour ces deux mesures. Lorsqu'un grand nombre d'images est considéré (plusieurs dizaines de milliers), une recherche exhaustive dont le temps de calcul est proportionnel au nombre d'images ($\mathcal{O}(N)$, où N est le nombre d'images) peut être un problème. Une solution est d'utiliser un algorithme de recherche de plus proche voisin approximatif, dont les temps de calcul sont généralement de l'ordre de $\mathcal{O}(\log(N))$. Dans notre implémentation, nous utilisons les algorithmes proposés par [Muja et Lowe 2009], dont le code est disponible publiquement et utilisable à partir de la libraire OpenCV. Ces travaux reposent entre autres sur l'utilisation de plusieurs arbres k-d, qui permettent chacun d'explorer très rapidement la base d'images, mais ne garantissent pas d'obtenir le voisin le plus proche dans tous les cas. La distance entre deux images est quantifiée par la distance euclidienne entre leurs caractéristiques. La valeur de cette distance n'intervient pas directement dans la formulation de l'algorithme Relief (si ce n'est dans le calcul du plus proche voisin), mais aura une importance dans les améliorations décrites en 2.3.3.

Distance entre les caractéristiques, fonction D_F

Pour quantifier la distance entre deux valeurs d'une caractéristique, nous avons jusqu'à présent évoqué la différence en valeur absolue entre ces valeurs. Cette mesure n'est pas du tout pertinente dans le cas de caractéristiques discrètes. Pour remédier à ce problème, dans le cas de variables discrètes, nous définissons la distance entre deux valeurs par $D_F(f_1,f_2) = 0$ si $f1 = f2$, et $D_F(f_1,f_2) = 1$ sinon. Si la mesure de distance est cette fois adaptée aux variables discrètes, le poids $\mathcal{R}(\mathcal{F})$ des caractéristiques discrètes a de fortes chances d'être surévalué par rapport aux données continues. En effet la distance entre deux valeurs sera souvent égale à 1, tandis que pour des variables continues, la distance sera généralement proche de l'écart-type de la variable, généralement largement inférieur à 1, les données étant réparties dans l'intervalle [0,1]. Pour compenser ce phénomène, nous utilisons la fonction rampe définie ci-dessous pour les variables continues :

$$D_F(f_1,f_2) = \begin{cases} 0 & \text{si } |f_1 - f_2| < t_{\min} \\ 1 & \text{si } |f_1 - f_2| > t_{\max} \\ |f_1 - f_2| & \text{sinon} \end{cases} \tag{2.2}$$

Les paramètres $t_{\min}$ et $t_{\max}$ sont à définir. La valeur $t_{\min}$ permet de ne pas tenir compte des très petites différences entre deux valeurs, qui ne fournissent généralement pas d'information. Nous posons $t_{\min} = 0{,}05$, en considérant que deux valeurs dont la différence absolue ne dépasse pas 0,05 ne sont pas informatives (présence de bruit, impossibilité de voir une telle différence à l'œil nu). A l'inverse, nous souhaitons que pour des différences élevées (de l'ordre de l'écart-type de la variable), la mesure de distance soit maximale. Nous choisissons donc de fixer $t_{\max} = 0{,}20$ car l'écart-type de nos variables prend des valeurs proches de 0,20. Ainsi, une différence de valeurs supérieure à 0,20, pour une variable continue $\mathcal{F}$, aura le même impact

sur le poids $\mathcal{R}(\mathcal{F})$ qu'une variable discrète dont les valeurs mesurées sont différentes.

Données manquantes

Il n'est pas possible de calculer de différences entre les caractéristiques lorsqu'il manque des informations concernant une image (valeurs notées -1, voir 2.2.1). Dans nos travaux nous considérons l'absence de données comme une information car celle-ci provient généralement de difficultés pour les algorithmes à évaluer correctement une valeur (mauvaise détection, obstruction, image de trop mauvaise qualité, etc.). Ainsi lorsqu'une valeur est comparée à une valeur correspondant à une donnée manquante, nous considérons que les deux valeurs n'ont rien en commun et pour toute valeur $f \neq -1$, $D_F(f, -1) = 1$. Nous fixons aussi $D_F(-1, -1) = 0$. Les données manquantes sont très rares dans nos jeux de données, toutefois il est important de garder à l'esprit que les algorithmes que nous utilisons présentent des limites et nous souhaitons également pouvoir évaluer les impressions véhiculées par une photo même lorsque toutes les données ne peuvent être calculées.

Nombre d'images parcourues

L'algorithme Relief est itératif : des images de la base d'apprentissage sont sélectionnées aléatoirement, leurs plus proches voisins dans chaque classe sont déterminés, et le poids de l'algorithme est mis à jour en tenant compte des différences entre les caractéristiques. Le nombre d'images sélectionnées est un paramètre important car il est directement proportionnel au temps d'exécution de l'algorithme, et influe fortement sur la pertinence du résultat. Dans nos travaux, étant donné que nous n'avons qu'un nombre restreint d'échantillons très bruités (subjectivité des scores, caractéristiques parfois peu pertinentes), nous faisons le choix d'utiliser l'ensemble de la base d'apprentissage plutôt qu'un échantillon aléatoire.

Formule

Finalement, nous pouvons résumer l'algorithme par l'équation 2.3 qui définit la valeur du poids $\mathcal{R}(\mathcal{F})$ à la fin du parcours des images. Nous notons N_a le nombre d'images de la base d'apprentissage, $D_F(.)$ la mesure de distance sur les caractéristiques. c_i représente la classe correspondant à l'image i et $k \in c_i$ est l'image la plus proche (au sens de la distance entre images définie précédemment) de l'image i appartenant à la même classe que i. $k \notin c_i$ est l'image la plus proche appartenant à l'autre classe. Enfin, $f_{k \notin c_i}$ et $f_{k \in c_i}$ sont respectivement les valeurs de la caractéristique $\mathcal{F}$ pour les images i, $k \notin c_i$ et $k \in c_i$.

$$\mathcal{R}(\mathcal{F}) = \frac{1}{N_a} \sum_{i=1}^{N_a} \left[D_F(f_i, f_{k \in c_i}) - D_F(f_i, f_{k \notin c_i}) \right] \tag{2.3}$$

2.3.3 Algorithme ReliefF - Classification généralisée

Principe

L'algorithme Relief défini précédemment est limité aux problèmes de classification binaires. Dans le cas de problèmes à plus de deux classes, la méthode doit être adaptée. La variante que nous détaillons ici est proposée initialement par [Kononenko 1994] et nommée ReliefF. L'algorithme amélioré intègre également d'autres paramètres permettant une meilleure gestion des données bruitées. L'équation 2.3 est adaptée et nous obtenons l'équation 2.4. En plus des notations utilisées précédemment, nous définissons le nombre de classes C, numérotées de 0 à $C-1$. Dans cette nouvelle version de l'algorithme, la distance entre les images est également considérée. Cette distance est prise en compte par la fonction $D_I(.)$, que nous définissons dans les paragraphes suivants. Un nouveau paramètre est également introduit : au lieu de considérer uniquement le plus proche voisin, pour chaque classe, ce sont les K plus proches voisins qui sont retenus et utilisés pour mettre à jour le poids de chaque variable. Ainsi, k_c^i désigne le $k^{\text{ème}}$ plus proche voisin de l'image i parmi les images de la classe c. Enfin, nous notons N_c le nombre d'images de la classe $c \in [\![0,C-1]\!]$.

$$\mathcal{R}(\mathcal{F}) = \sum_{i=1}^{N_a} \left[\underbrace{\left(\sum_{c \neq c_i}^{C} \underbrace{\frac{N_c}{N_a - N_{c_i}}}_{A} \frac{\sum_{k=1}^{K} D_F(f_i, f_{k_c^i}) D_I(i, k_c^i)}{\sum_{k=1}^{K} D_I(i, k_c^i)} \right)}_{B} - \underbrace{\frac{\sum_{k=1}^{K} D_F(f_i, f_{k_{c_i}^i}) D_I(i, k_{c_i}^i)}{\sum_{k=1}^{K} D_I(i, k_{c_i}^i)}}_{C} \right] \tag{2.4}$$

Nombre de voisins considérés K

Un premier paramètre important à définir dans cette nouvelle équation est le nombre de voisins considérés. Celui-ci a un impact sur le temps de calcul, mais surtout sur la robustesse de l'algorithme. En effet dans le cas où un seul voisin est étudié, celui-ci peut être mal évalué ou posséder des caractéristiques extrêmes peu représentatives du cas général. Il est ainsi intéressant de moyenner les résultats à partir de plusieurs observations. Les travaux de [Robnik-Šikonja et Kononenko 2003] montrent qu'au-delà de 70 voisins, les performances de l'algorithme montrent leurs limites. [Robnik-Šikonja et Kononenko 2003] indiquent également que pour la plupart des problèmes, un nombre plus réduit de l'ordre de 10 voisins explorés est largement suffisant. Nos premières expériences nous ont montré que fixer $K = 10$ permet d'obtenir des résultats pertinents sur différentes bases d'images, et nous conservons donc cette valeur dans la suite de ce document.

Distance entre images, fonction $D_I(.)$

Dans l'équation 2.4, le terme $D_I(i_1,i_2)$ est une fonction de distance entre les images i_1 et i_2. L'objectif est de valoriser les voisins les plus proches de l'image cible au détriment des voisins plus éloignés, car les images plus proches sont de meilleurs indicateurs. En général, une décroissance exponentielle par rapport à la distance est proposée. En notant $D_E(i,j)$ la distance euclidienne entre les images i et j dans l'espace des caractéristiques, une première possibilité est donc :

$$D_I(i,j) = \frac{e^{-\frac{D_E(i,j)}{\sigma}^2}}{\sum_{k=1}^{K} D_E(i,k)} \tag{2.5}$$

La distance entre une image et son voisin est normalisée par la somme des différences entre l'image et ses voisins d'une classe donnée (images $k \in [\![1,K]\!]$). Cette expression est toutefois très dépendante des données, et le poids du premier voisin ne sera pas le même pour chaque image. Ainsi [Robnik-Šikonja et Kononenko 2003] proposent de remplacer la distance réelle entre les images ($D_E(.)$) par le rang de l'image (1 pour le voisin le plus proche, 2 pour le second, etc.). La fonction utilisée dans notre implémentation est donc la suivante :

$$D_I(i,j) = \frac{e^{-\frac{rank(i,j)}{\sigma}^2}}{\sum_{k=1}^{K} D_E(i,k)} \tag{2.6}$$

où $rank(i,j)$ désigne le nombre d'images plus proches de l'image i que son voisin j, pour une classe fixée. Nous fixons $\sigma = K-1$ de manière à ce que le dernier voisin considéré ait un poids inférieur à celui du plus proche voisin, sans pour autant être négligeable :

$$\text{Pour K} = 10, \frac{e^{-\frac{1}{K-1}^2}}{e^{-\frac{K}{K-1}^2}} \approx 0{,}33 \tag{2.7}$$

Chacun des K voisins aura donc un impact sur le poids global de la caractéristique, le dernier voisin ayant un poids 3 fois moins élevé que le plus proche voisin. Cela permet d'obtenir un résultat plus robuste au bruit par rapport au cas où un seul voisin est considéré.

Termes de l'équation

L'équation 2.4 donne le détail des calculs permettant d'arriver à une estimation de la pertinence d'une variable. Par rapport à l'équation 2.3, trois modifications majeures sont effectuées. La première consiste en la prise en compte d'un nombre quelconque de classes dans l'algorithme : pour une image donnée, les plus proches voisins sont calculés dans chaque classe. Le terme de droite (C) de l'équation correspond au second terme de l'équation 2.3, avec les changements suivants :

— Plusieurs (K) plus proches voisins sont pris en compte,

— L'importance de chaque voisin est pondérée par la fonction $D_I(.)$.

Ces changements permettent d'accroître la robustesse de l'algorithme au bruit (plusieurs voisins sont considérés) en prenant en compte plus d'informations. L'expression B de l'équation correspond au premier terme de l'équation 2.3 avec les mêmes changements que le second terme. Enfin, le terme A permet de normaliser le poids de chaque classe : les classes contenant beaucoup d'images (N_c élevé) ont plus de poids. De plus, pour chaque image, la contribution de la classe de l'image est la même que la somme des contributions des autres classes.

Amélioration proposée

Le terme A de l'équation 2.4 permet de prendre en compte l'importance de chaque classe en fonction du nombre d'images contenues dans la classe. Toutefois, dans nos travaux, les classes sont définies selon des niveaux particuliers : photos très peu (classe 0) / peu (classe 1) / adaptées (classe 2) / très adaptées (classe 3) à une application donnée. Il existe donc une sorte de hiérarchie entre les classes et les images de très mauvaise qualité sont donc plus éloignées des images de très bonne qualité. Pour intégrer ce facteur dans nos expressions, nous modifions le terme A de l'équation afin d'ajouter une information concernant l'écart entre la classe de l'image étudiée et celle de son voisin. Nous proposons ainsi :

$$A = \frac{N_c \times |c - c_i|}{N_a - N_{c_i}} \tag{2.8}$$

Afin de renormaliser le rapport A, nous divisons ce terme par la somme :

$$\sum_{c \neq c_i}^{C} |c - c_i| = \sum_{c=0}^{c_i} (c - c_i) + \sum_{c=c_i}^{C-1} (c_i - c) = 0{,}5(C^2 - C - 2c_iC + 2c_i + 2c_i^2) \tag{2.9}$$

Finalement,

$$A = \frac{2N_c|c - c_i|}{(N_a - N_{c_i})(C^2 - C - 2c_iC + 2c_i + 2c_i^2)} \tag{2.10}$$

La définition de ce nouveau poids permet d'adapter l'algorithme ReliefF à notre problématique, dans laquelle les classes correspondent à différents niveaux de pertinence des photographies.

Remarque : Dans le cas de la classification binaire, nous choisissons également d'utiliser les formules définies dans cette partie, notamment afin d'exploiter les informations de plusieurs voisins. D'ailleurs, lorsqu'il n'y a que deux classes, les équations 2.3 et 2.4 ne diffèrent que par l'introduction des informations liées aux K plus proches voisins, le terme B devient une somme d'un seul terme et le rapport A est égal à 1. La formule, dans le cas où $C = 2$ se simplifie donc ainsi :

$$\mathcal{R}(\mathcal{F})_{C=2} = \sum_{i=1}^{N_a} \left[\frac{\sum_{k=1}^{K} D_F(f_i, f_{k^i_{1-c_i}}) D_I(i, k^i_{1-c_i})}{\sum_{k=1}^{K} D_I(i, k^i_{1-c_i})} - \frac{\sum_{k=1}^{K} D_F(f_i, f_{k^i_{c_i}}) D_I(i, k^i_{c_i})}{\sum_{k=1}^{K} D_I(i, k^i_{c_i})} \right] \tag{2.11}$$

2.3.4 Algorithme RReliefF - Régression

Dans le cas de la régression, il est impossible de calculer les plus proches voisins pour chaque classe, car les images sont associées à un score. Une possibilité est d'ordonner les images par scores et de recréer des catégories afin de retomber sur le problème précédent. Toutefois

les petites différences entre les scores sont alors supprimées et de l'information est perdue. La méthode que nous proposons est directement inspirée des travaux de [Robnik-Šikonja et Kononenko 1997], dont l'idée repose sur le principe suivant. Une caractéristique $\mathcal{F}$ doit avoir une valeur de $\mathcal{R}(\mathcal{F})$ élevée lorsque pour deux images similaires (au sens des caractéristiques) dont les évaluations sont différentes (score différent), leur valeur de $\mathcal{F}$ est également différente. A l'inverse, lorsque deux images similaires ont des évaluations proches, leur valeur de $\mathcal{F}$ est également proche. Ce principe peut se formaliser de la façon suivante :

$$\begin{aligned}\mathcal{R}(\mathcal{F}) =& \text{Prob(Différentes valeurs de } \mathcal{F} \mid \text{Images voisines et Évaluations différentes)} \\ &- \text{Prob(Différentes valeurs de } \mathcal{F} \mid \text{Images voisines et Évaluations identiques)}\end{aligned} \tag{2.12}$$

Ce principe est le même que celui utilisé dans le cas de la classification : chacun des deux termes de l'équation 2.12 correspond respectivement aux termes B et C de l'équation 2.4. La difficulté consiste à évaluer ce que nous considérons comme des évaluations différentes ou identiques. Pour cela, nous normalisons dans un premier temps les scores de la même manière que les caractéristiques : le plus petit score est fixé à 0, le plus grand à 1. Deux scores sont considérés comme différents lorsque leur différence est proche de 1, et proches lorsqu'elle est proche de 0. En notant S_i le score normalisé de l'image i, nous définissons la fonction de distance entre les scores de deux images i et j par $D_S(i,j) = |S_i - S_j|$. Pour chaque image i, nous calculons ses K plus proches voisins et en reprenant l'équation 2.11, nous multiplions le premier terme par D_S (qui représente la différence entre les évaluations), et le second terme par $1 - D_S$ (similarité entre les évaluations). Nous obtenons finalement :

$$\mathcal{R}(\mathcal{F}) = \sum_{i=1}^{N_a} \sum_{k=1}^{K} \left[\frac{D_F(f_i,f_{k^i}) D_I(i,k^i) D_S(i,k^i)}{\sum_{k=1}^{K} D_I(i,k^i)} - \frac{D_F(f_i,f_{k^i}) D_I(i,k^i)(1 - D_S(i,k^i))}{\sum_{k=1}^{K} D_I(i,k^i)} \right] \tag{2.13}$$

Finalement, les caractéristiques les plus discriminantes sont celles dont les valeurs $\mathcal{R}(\mathcal{F})$ sont les plus grandes. En triant les caractéristiques par ordre d'importance, il est ainsi possible de n'utiliser que des caractéristiques informatives dans nos évaluations. Il est également possible de supprimer automatiquement des caractéristiques trop peu discriminantes, en ne conservant par exemple que celles dont les valeurs $\mathcal{R}(\mathcal{F})$ sont strictement positives.

2.4 Apprentissage automatique et prédiction

2.4.1 Algorithmes d'apprentissage

Il existe de nombreux algorithmes d'apprentissage automatique permettant la prédiction de variables, continues ou discrètes. Nous avons retenu 4 d'entre eux : les machines à vecteur de support [Vapnik et Vapnik 1998] (SVM, pour Support Vector Machines), les réseaux de

neurones artificiels [Riedmiller et Braun 1993 ; LeCun et al. 1998] (ANN, pour Artificial Neural Networks), les forêts aléatoires [Breiman 2001] (RF, pour Random Forest) et les forêts boostées [Friedman 2001] (GBT, pour Gradient Boosted Trees). Nous avons retenu ces algorithmes car :

— Ils peuvent s'adapter aux différents problèmes (classification, régression) et types de caractéristiques (continues, discrètes).
— Ils ne fonctionnent pas sur le même principe, il est donc possible que certains soient plus adaptés à des données particulières.
— Ils sont déjà implémentés et correctement documentés dans la bibliothèque OpenCV.

D'autres méthodes existent. Nous avons par exemple testé l'algorithme des k plus proches voisins, qui consiste à regarder les catégories ou les scores des images les plus proches de l'image cible dans l'espace des caractéristiques, et à évaluer la cible en fonction des évaluations des voisins. Les résultats obtenus sont cependant quasiment toujours en-dessous de ceux obtenus pour chacun des autres algorithmes présentés et nous ne l'utilisons donc pas dans nos modèles. Actuellement, de très nombreux travaux reposent sur l'utilisation de réseaux de neurones convolutionnels ou CNN [LeCun et al. 1998]. Toutefois, ce type d'algorithme nécessite un très grand nombre de données pour l'apprentissage ce qui empêche leur utilisation dans notre cas.

Dans cette partie, nous décrivons brièvement le principe de fonctionnement de chacun des algorithmes que nous utilisons et détaillons les différents paramètres que nous avons à fixer lors de leur utilisation. Dans les travaux présentés dans ce document, nous fixons généralement ces valeurs afin d'étudier en priorité l'influence des caractéristiques sur les performances et de pouvoir comparer des modèles créés à partir de paramètres identiques.

2.4.1.1 Machines à vecteur de support

Les machines à vecteurs de support (ou Support Vector Machines, SVM) sont largement utilisées pour résoudre les problèmes de classification. Elles ont été initialement conçues pour distinguer deux classes d'objets, puis la méthode a été étendue à la classification à un nombre quelconque de classes ainsi qu'à la régression [Drucker et al. 1997].

Principe

Nous notons N_C le nombre total de classes et N_F le nombre de caractéristiques décrivant chaque image. L'algorithme cherche à déterminer les séparations optimales dans l'espace des caractéristiques de manière à ce que chaque région définie par ces séparations contienne uniquement des images d'une classe particulière. Ce principe est résumé sur la figure 2.4 dans le cas où $N_F = 3$ et $N_C = 2$. En pratique, après apprentissage il suffit de regarder où se trouvent les nouvelles images par rapport à l'hyperplan P (cf. figure 2.4) dans cet espace afin de déterminer leur catégorie. Dans le cas général, les régions ne sont pas linéairement séparables et la surface de séparation optimale n'est pas un plan. Des méthodes existent pour calculer de telles surfaces, en projetant les points dans un espace de plus grande dimension (astuce du noyau). Des techniques permettent également de traiter les cas où $N_C > 2$, en calculant

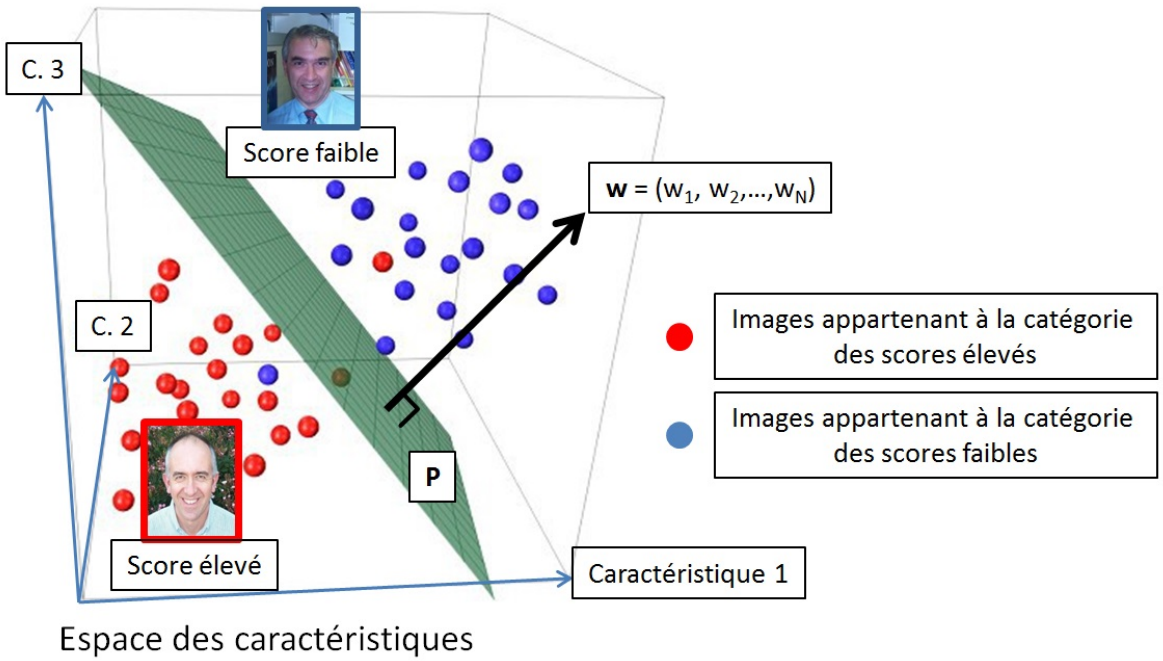


FIGURE 2.4 – Machines à vecteur de support. Exemple de séparation pour 2 classes d'images, représentées dans un espace de caractéristiques à 3 dimensions. La surface de séparation est ici un plan car les données sont (à deux images près) linéairement séparables.

par exemple des plans séparant les classes 2 à 2 ($\frac{N_C(N_C-1)}{2}$ plans sont ainsi nécessaires). Il est également possible d'adapter la méthode afin de faire de la régression. Pour cela, le problème d'optimisation est modifié et intègre des contraintes sur la distance entre le score de l'image et sa projection sur l'hyperplan séparateur.

La bibliothèque OpenCV intègre toutes les fonctionnalités issues de la bibliothèque libSVM dont la documentation et le détail du fonctionnement sont explicités dans l'article de [Chang et Lin 2011]. Les paragraphes suivants présentent les différents paramètres et variantes de l'algorithme.

Contraintes d'optimisation

Selon le problème (classification ou régression), les contraintes d'optimisation ne sont pas les mêmes. De plus, pour un même problème, différentes variantes de SVM existent. Dans le cas de la classification (nous notons SVC pour Support Vector Classification), les variantes sont :

— C-SVC : C est un paramètre introduit dans le problème d'optimisation de manière à gérer les données aberrantes. Ainsi, plutôt que de tenter de trouver une séparation parfaite entre les données (risque de surapprentissage des données aberrantes), une séparation imparfaite peut être proposée par l'algorithme. Une pénalité proportionnelle à C est alors introduite pour les données mal classées.

— ν-SVC : ν permet également de traiter les données aberrantes. Une valeur de ν proche de 0 impose des contraintes sur les vecteurs de support afin d'obtenir très peu d'erreurs de classification sur la base d'apprentissage, le risque étant de perdre en capacité de généralisation (surapprentissage). Ainsi, la surface de séparation est d'autant plus lisse que la valeur de ν est grande, et d'autant plus précise que la valeur de ν est petite.

Dans le cas de la régression (nous notons SVR pour Support Vector Regression), les variantes sont :

— ϵ-SVR : ϵ est la distance maximale acceptable entre le score de l'image et sa projection sur l'hyperplan séparateur. Les images au-delà de cette distance introduisent une pénalité dans la valeur à minimiser. Cette pénalité peut être ajustée à l'aide du paramètre C (voir C-SVC).

— ν-SVR : Le principe de ν-SVR est le même que pour ν-SVC. Le paramètre C permet de trouver un compromis entre généralisation et précision.

Pour un jeu de données particulier, deux variantes différentes peuvent produire en général des résultats différents. Toutefois nos essais sur différentes bases d'images et différentes caractéristiques nous ont montré qu'il n'y a pas de variante plus performante qu'une autre dans le cas général. Nous choisissons d'utiliser C-SVC pour la classification, et ν-SVR pour la régression pour nos expériences dans tout le reste de ce document. Ce choix est arbitraire et nous fixons ces paramètres dans nos expériences, dans lesquelles nous n'étudierons que l'influence des caractéristiques et des bases d'images sur les performances.

Paramètres C, ν et ϵ

Ces paramètres permettent d'ajuster le poids des données mal classées dans les différentes variantes du problème d'optimisation. Leurs valeurs doivent être strictement positives et dans le cas de ν et ϵ, inférieures à 1. C est fixé par défaut à 1 et les paramètres ν et ϵ à 0 dans la bibliothèque OpenCV. Comme ν et ϵ doivent être strictement positifs, nous fixons leur valeur à 0,1 dans tous nos tests, pour la classification et la régression. Il est possible d'optimiser ces valeurs par validation croisée sur la base d'apprentissage. Cette dernière étape n'améliore pas les performances de nos modèles, et ralentit considérablement le temps d'apprentissage. Toutefois, dans le cas où les caractéristiques considérées sont de différentes natures (certaines continues, d'autres discrètes), nous procédons à cette étape d'optimisation afin d'améliorer la stabilité des résultats.

SVM non linéaire : choix de la fonction noyau

Les fonctions noyaux permettent de projeter l'espace des caractéristiques dans un espace de plus grande dimension (éventuellement infini) afin de rendre possible une séparation linéaire des données dans ce nouvel espace. Une projection inverse permet alors d'obtenir les vecteurs de support dans l'espace des caractéristiques. Cette étape permet d'introduire une séparation non linéaire entre les données. L'intérêt d'utiliser une fonction noyau est d'autant plus grand qu'il n'est pas nécessaire d'avoir à faire des calculs dans cet espace de grande dimension pour résoudre le problème. Seule l'expression du produit scalaire entre deux vecteurs de l'espace de

caractéristiques est à définir. LibSVM propose de choisir parmi un noyau linéaire, polynomial, sigmoïdal (tangente hyperbolique) et gaussien. La plupart des travaux utilisant des SVM tendent à montrer qu'un noyau gaussien est généralement efficace. Pour deux vecteurs x_i et x_j, l'expression du produit scalaire $\mathcal{K}$ correspondant au noyau gaussien s'écrit $\mathcal{K}(x_i,x_j) = exp(\gamma||x_i - x_j||^2)$. Ce noyau est également le noyau proposé par défaut dans la bibliothèque OpenCV ; nous utilisons ainsi dans ce document un noyau gaussien dont le paramètre γ est égal à 1. Nous n'avons pas observé de différence significative dans nos résultats en faisant varier le paramètre γ, et celui-ci peut être optimisé par validation croisée sur la base d'apprentissage.

Critères d'arrêt de l'apprentissage

L'algorithme d'optimisation s'arrête lorsque le nombre maximal d'itérations autorisé est atteint. Par défaut, la valeur fixée par OpenCV est de 1000 itérations, et nous conservons cette valeur car nous ne constatons pas de changements significatifs lorsque cette valeur est modifiée.

2.4.1.2 Réseaux de neurones artificiels

Les réseaux de neurones artificiels (ANN) peuvent être utilisés afin de faire de la prédiction de classes ou de scores. Concrètement, un ensemble de cellules (les neurones) sont connectées entre elles à la manière d'un automate transformant des entrées en sortie(s) selon des règles précises.

Principe

Le réseau utilisé dans notre programme est un réseau où les informations sont propagées uniquement vers l'avant du réseau ; il n'y a pas de boucle dans le réseau. Le principe de fonctionnement de l'algorithme d'apprentissage est présenté sur la figure 2.5. Les zones oranges (claires) entourent les paramètres appris lors de l'apprentissage, les zones rouges (foncées) ceux qui ont besoin d'être connus dès le départ. Les valeurs des caractéristiques extraites de l'image sont représentées en vert (à gauche du réseau) et les scores des images en jaune (à droite du réseau). Dans le cas de la classification, ces scores sont des étiquettes représentées par des valeurs entières. L'ensemble des neurones présents dans la couche cachée (en bleu) assure la flexibilité du modèle.

Différents algorithmes permettent d'optimiser les paramètres afin d'obtenir les prédictions les plus proches des scores attendus. Les algorithmes utilisent généralement le principe proposé par [Rumelhart et al. 1988] de la propagation des erreurs vers l'arrière (entrée du réseau). Une heuristique améliorant la rapidité de convergence de l'algorithme a été développée par [Riedmiller et Braun 1993] et se nomme "Rprop", pour "Resilient Backpropagation". Cette technique est utilisée dans nos expériences car c'est une méthode rapide de mise à jour des coefficients, adaptée à notre type de réseau où les informations sont propagées uniquement vers l'avant (sortie) du réseau.

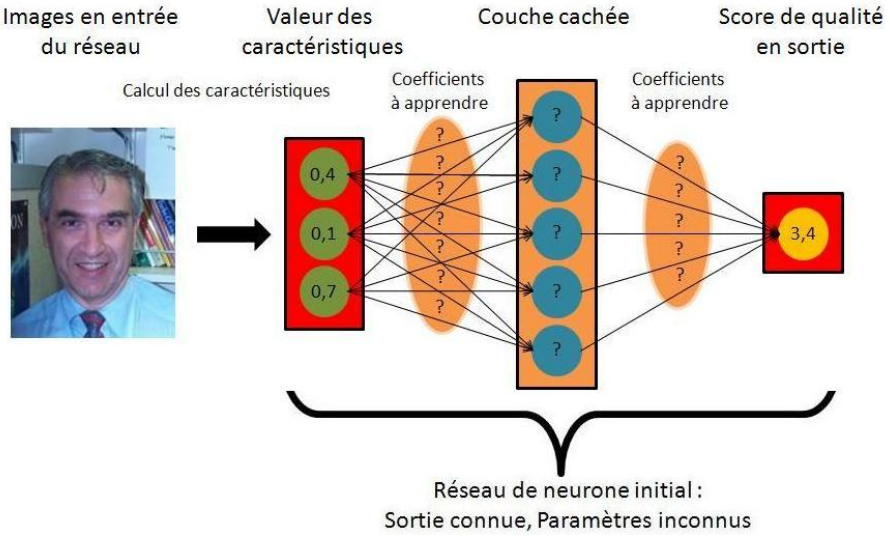


FIGURE 2.5 – Apprentissage à l'aide d'un réseau de neurones. Les coefficients à apprendre sont les connexions entre les différents neurones.

Une fois les coefficients estimés, il est possible d'ajouter de nouvelles photos en entrée et de calculer automatiquement un score en sortie, comme indiqué sur la figure 2.6.

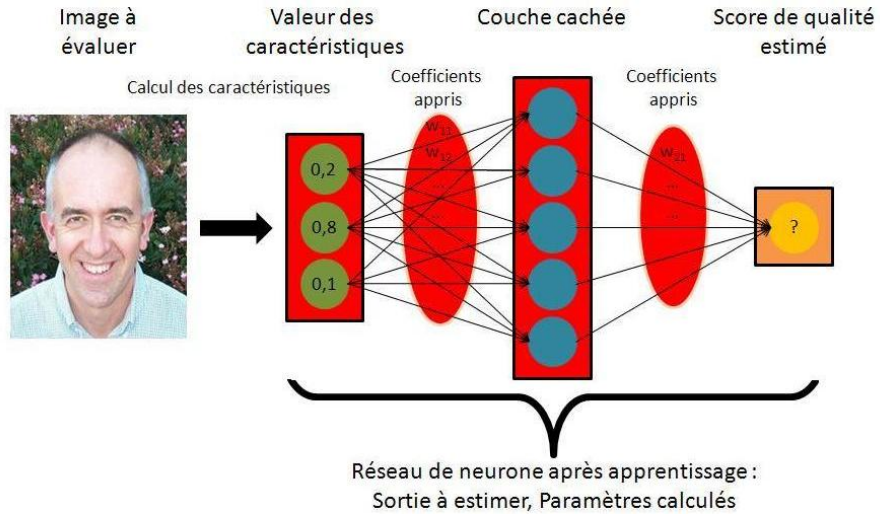


FIGURE 2.6 – Prédiction à l'aide d'un réseau de neurones.

Structure du réseau

La structure d'un réseau de neurones peut être complexe, et surtout différente d'un problème à l'autre. Le premier paramètre à définir est le nombre de couches cachées (une seule sur les figures 2.5 et 2.6) dans le réseau. La plupart des réseaux traditionnels ne considèrent qu'une seule couche, car les résultats ont tendance à ne pas s'améliorer lorsque la complexité du réseau

augmente. Récemment, de plus en plus de travaux utilisent des réseaux profonds contenant plusieurs couches cachées, cependant nos jeux de données étant réduits, nous ne considérons qu'une seule couche afin d'éviter les problèmes de surapprentissage liés à un trop grand nombre de paramètres dans le réseau. Un autre paramètre correspond au nombre de neurones de la couche cachée. Nous avons choisi de fixer ce nombre en fonction du nombre de caractéristiques d'entrées car plus ce nombre est élevé, plus la dépendance entre les caractéristiques peut être forte et il peut être nécessaire d'avoir à combiner un grand nombre d'informations dans les neurones de la couche intermédiaire. Si N_F est le nombre de caractéristiques, nous fixons le nombre de neurones cachés à $N_F/2$. Cette valeur est souvent utilisée dans les travaux utilisant des réseaux de neurones, et nous obtenons ainsi des résultats satisfaisants quel que soit le nombre de caractéristiques en entrée. Nous avons également constaté qu'augmenter le nombre de neurones au-delà d'un certain seuil n'augmente plus les performances, nous limitons ainsi le nombre de neurones de la couche cachée à 20. Cela peut sembler peu en comparaison des réseaux utilisés actuellement pour la classification d'images ou la détection d'objets, toutefois il est difficile de créer des réseaux plus complexes alors que nous n'avons que peu d'images disponibles pour l'apprentissage. Limiter le nombre de neurones réduit également le risque de surapprentissage car il y a moins de paramètres dans le réseau.

Fonction d'activation

Les coefficients de chaque connexion sont multipliés aux sorties des neurones, et la somme de ces produits est fournie en entrée des neurones de la couche suivante. Les neurones sont ensuite activés en fonction de la somme de ces contributions : si la valeur d'entrée est faible, le neurone n'est pas activé (sortie égale à -1). Si elle est élevée, la sortie du neurone vaut 1 (le neurone est activé). L'activation ou non des neurones est ajusté à l'aide d'une fonction d'activation, généralement modélisée par une courbe sigmoïdale de la forme $f(x) = \beta * (1 - e^{-\alpha x})/(1 + e^{-\alpha x})$ où x est la somme des contributions des neurones de la couche inférieure. OpenCV propose des valeurs par défaut de 1 pour α et β. Nous conservons ces valeurs dans nos expériences.

Critères d'arrêt de l'apprentissage

L'algorithme d'apprentissage des poids optimaux, Rprop, s'arrête lorsque les données sont correctement classées. Dans le cas de la régression, l'algorithme s'arrête également lorsque les prédictions sont très proches des scores de vérité terrain. L'algorithme continue ainsi d'ajuster les poids jusqu'à ce que l'un des deux critères suivants soient remplis :

— Le nombre maximal d'itérations autorisé est atteint.

— L'erreur de prédiction se situe en-dessous d'une certaine valeur.

Par défaut, les valeurs fixées par OpenCV sont de 100 pour le nombre maximal d'itérations possibles, et de 0,01 pour l'erreur de prédiction moyenne sur l'ensemble de la base d'apprentissage. Nous avons constaté que dans le cas de la régression, diminuer l'erreur de prédiction autorisée à 0,001 améliore les performances. Une explication possible est le faible écart-type des scores dans les jeux de données dont nous disposons : il est ainsi important de s'approcher

au mieux des scores de vérité terrain lors de l'apprentissage. Afin d'atteindre cette faible erreur de prédiction, nous augmentons également le nombre d'itérations maximales possibles à 1000. Cela ne constitue pas un problème concernant le temps de calcul car le nombre d'itérations est compensé par le faible nombre d'images.

2.4.1.3 Forêts aléatoires

Les forêts aléatoires (notées RF pour Random Forest) reposent sur l'utilisation de différents arbres de décision générés aléatoirement. Les scores ou les classes prédites par chaque arbre de décision sont moyennés afin de faire une prédiction globale robuste. La méthode a été proposée par [Breiman 2001].

Principe

Le principe d'un arbre de décision, décrit dans les travaux de [Breiman et al. 1984], est résumé sur la figure 2.7. Concrètement, chaque nœud de l'arbre correspond à une prise de décision concernant une caractéristique de l'image. Les feuilles de l'arbre correspondent à une décision globale sur l'image (score, étiquette de classe). Lors de la phase de prédiction, il suffit ainsi de parcourir chaque arbre afin de regarder les prédictions de la feuille correspondant à l'image. Dans le cas de la classification, la prédiction finale correspond à la classe obtenant le plus de votes parmi l'ensemble des arbres. Dans le cas de la régression, la prédiction finale est la moyenne de la prédiction de chaque arbre. Notons que le fait de moyenner un ensemble de prédictions produit généralement une variance plus faible des résultats : l'erreur moyenne de prédiction sera potentiellement plus faible mais les erreurs correspondant aux scores extrêmes seront souvent supérieures à celles obtenues par SVM ou ANN.

Construction de la forêt

Pour chaque arbre, un sous-ensemble choisi aléatoirement des images de la base d'apprentissage est utilisé. De même, seul un sous-ensemble choisi aléatoirement dans l'ensemble des caractéristiques est considéré afin de procéder à la séparation des branches de l'arbre. Ce double échantillonnage rend le modèle général très robuste au bruit et peu probable le surapprentissage. Généralement le nombre d'images utilisé pour l'apprentissage de chaque arbre est de l'ordre de $2N_a/3$, où N_a est le nombre d'images de la base d'apprentissage. En notant N_F le nombre de caractéristiques, $\sqrt{N_F}$ caractéristiques sont utilisées pour le calcul de chaque nœud de chaque arbre. Utiliser plus de caractéristiques et/ou d'images pour chaque arbre augmente les risques de surapprentissage. Nous conservons ces valeurs dans nos expériences. Dans ce document, nous ne détaillons pas le processus de construction des arbres de décision, et nous évoquons simplement les différents paramètres permettant d'adapter les forêts d'arbres aléatoires à un problème donné.

Paramètres de construction des arbres

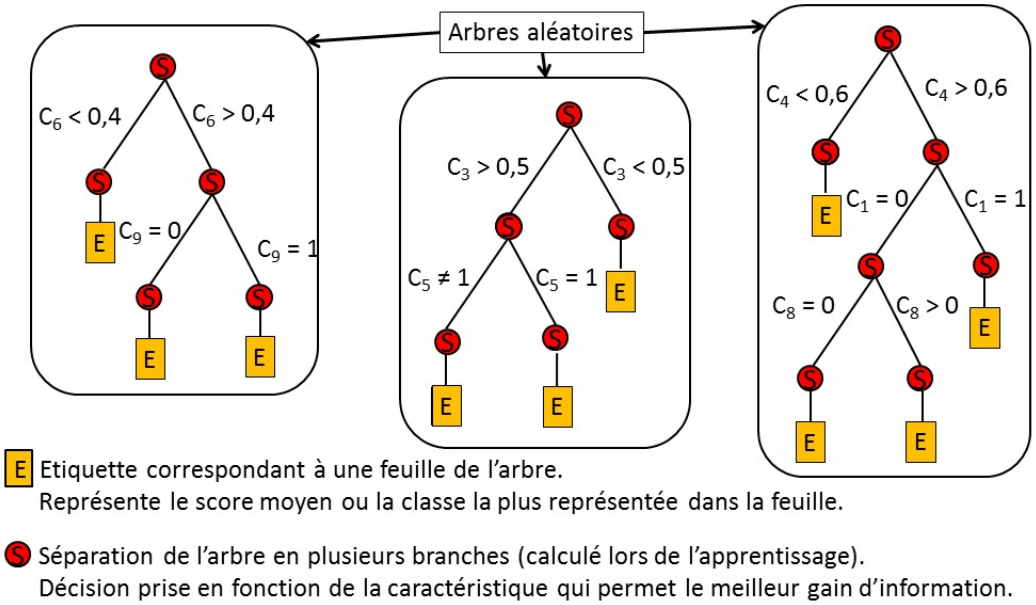


FIGURE 2.7 – Exemple représentant 3 arbres de décision obtenus à partir d'un jeu de données. Pour la construction d'une forêt aléatoire, chaque arbre est élaboré à partir d'un échantillonnage de la base d'apprentissage. Chaque nœud de chaque arbre est optimisé à partir d'un échantillonnage de l'espace des caractéristiques.

Différents paramètres permettent de contrôler la construction de chaque arbre. En effet, si les arbres sont trop profonds, des problèmes de surapprentissage peuvent apparaître. Ainsi, en plus de n'utiliser qu'un nombre restreint d'images et de caractéristiques, la profondeur de chaque arbre est limitée. Nous conservons la valeur 5 proposée par défaut par la bibliothèque OpenCV comme limite de profondeur. Une autre valeur contrôlant la construction des arbres est le nombre d'images requises pour qu'une feuille soit divisée. Ce nombre est fixé à 10 par défaut et nous conservons cette valeur. En effet, nous considérons qu'en-dessous de 10 images il n'existe pas assez d'information permettant de prendre une décision pour distinguer deux groupes d'images.

Critères d'arrêt de l'apprentissage

L'algorithme crée des arbres de décision jusqu'à ce qu'une des conditions suivantes soient remplies :

— Le nombre maximal d'arbres autorisé est atteint. Ce nombre est fixé à 50 par défaut et nous ne le modifions pas. Le temps dédié à l'apprentissage ainsi qu'à la prédiction est linéairement dépendant du nombre d'arbres construits. Notons que différents travaux s'intéressent actuellement à l'utilisation de ces algorithmes dans le cadre de la programmation parallèle : la construction et la prédiction de chaque arbre étant indépendantes, il est possible de paralléliser massivement les forêts aléatoires.

— L'erreur de prédiction de la forêt sur l'ensemble d'apprentissage est suffisamment faible.

Cette erreur est fixée à 0,1 par défaut et nous ne la modifions pas.

2.4.1.4 Forêts boostées

La différence entre les forêts aléatoires et les forêts boostées (GBT pour Gradient Boosted Trees) est le processus de création de la forêt. Dans le premier cas les arbres sont générés aléatoirement, tandis que les forêts boostées sont construites itérativement selon le principe du boosting. L'objectif du boosting est de réduire l'erreur de prédiction à chaque itération (une itération correspond à la création d'un arbre), selon le même principe que les algorithmes d'optimisation reposant sur une descente de gradient. La phase de prédiction est la même que dans le cas des forêts aléatoires : la prédiction de la forêt correspond à la combinaison des prédictions de chaque arbre.

Construction de la forêt

Tout d'abord, l'algorithme est initialisé en proposant le modèle constant le plus performant. Par exemple, la prédiction sera à chaque fois égale à la moyenne des scores des images de manière à réduire au maximum l'erreur quadratique moyenne. La forêt est alors construite itérativement en répétant M fois les étapes suivantes, détaillées dans les travaux de [Friedman 2001] :

1. Pour $t\%$ des images choisies aléatoirement dans la base d'apprentissage, l'erreur de prédiction est calculée. Ce taux est fixé par défaut à $t = 80\%$ dans OpenCV et peut être modifié, mais nous conservons cette valeur dans nos expériences. Ne pas utiliser toutes les images limite le risque de surapprentissage de la même manière que pour les forêts aléatoires.
2. Un nouvel arbre de décision est construit de façon à ce que la prédiction de cet arbre permette de corriger l'erreur de prédiction calculée à l'étape précédente.
3. Le modèle global est ajusté en intégrant ce nouvel arbre, en ajustant le poids de cet arbre de façon à minimiser l'erreur de prédiction globale de la forêt. Il est possible d'ajouter un paramètre de régularisation $\nu \in]0,1]$ afin de réduire le poids de chaque nouvel arbre afin d'éviter les situations de surapprentissage.

En notant y_i le score associé à chaque image x_i, et T_m l'arbre construit lors de l'itération m, le modèle final peut ainsi s'écrire de la façon suivante :

$$\text{Prediction}(x_i) = \text{Constante} + \nu \sum_{m=1}^{M} T_m(x_i) \tag{2.14}$$

Paramètres de construction des arbres

Un premier paramètre permettant de contrôler la construction des arbres est la fonction L définissant le calcul de l'erreur de prédiction utilisée pour le boosting. Nous conservons la

fonction proposée par défaut par la bibliothèque OpenCV, correspondant à l'erreur quadratique moyenne : $L(y,f(x)) = 0{,}5(y - f(x))^2$. Le paramètre de régularisation ν est fixé à 0,01, et nous constatons qu'augmenter cette valeur a tendance à diminuer les performances. Cette valeur est donc utilisée dans nos expériences. Enfin, de la même manière que pour les forêts aléatoires, la profondeur maximale des arbres ainsi que le nombre minimal d'images requises pour diviser une feuille peuvent également être modifiés. Nous conservons les valeurs par défaut, qui sont de 3 pour la profondeur maximale des arbres (5 pour les forêts aléatoires) et 10 pour le nombre d'images nécessaires à la division d'une feuille.

Critères d'arrêt de l'apprentissage

L'algorithme s'arrête lorsque M arbres de décision sont construits. Ce nombre est fixé à 200 dans l'implémentation proposée par OpenCV, ce qui est supérieur au nombre d'arbres construits par les forêts aléatoires (50). Ce nombre est élevé car le paramètre ν est faible : le nombre d'arbres nécessaires à la convergence de l'algorithme est donc élevé.

2.4.2 Protocole et critères d'évaluation

2.4.2.1 Validation croisée

Afin de réaliser l'apprentissage et la classification, la base de données est découpée en deux ensembles. Le premier est utilisé pour l'apprentissage tandis que le second permet de tester le modèle sur de nouvelles images. Le choix des ensembles d'apprentissage et de test est crucial dans la procédure d'évaluation des différents algorithmes proposés. Par exemple, utiliser trop peu d'images pour l'apprentissage ne permet pas de développer un modèle robuste capable de prédire correctement les scores ou classes des images utilisées pour les tests. A l'inverse, laisser trop peu d'images pour les tests rend les performances de prédiction peu significatives. Il faut également tenir compte de la représentation de chaque catégorie d'images dans chacune des deux bases. En effet, utiliser uniquement des photos de mauvaise qualité dans la base d'apprentissage ne permettra pas d'évaluer correctement des photos de bonne qualité lors des tests. Enfin, les bases d'apprentissage et de test doivent être distinctes afin de ne pas fausser les résultats.

Pour éviter au mieux ces problèmes, nous avons choisi l'approche de la validation croisée. L'idée consiste à diviser la base d'images en k échantillons distincts, puis à utiliser un échantillon pour les tests et les $k-1$ restants pour l'apprentissage. En répétant cette opération pour les k échantillons, nous obtenons une prédiction pour chaque image de la base. Pour des valeurs de k élevées (par exemple 10) cela permet d'utiliser 90% de la base pour l'apprentissage, ce qui suffit généralement à établir un modèle pertinent. L'inconvénient de cette approche est la nécessité de générer k modèles, et donc d'effectuer k apprentissages. Toutefois, nos jeux de données étant réduits, nous aurons tendance à privilégier une valeur de k élevée car le temps d'apprentissage de chaque modèle est suffisamment faible. Dans tous les tests effectués dans ce document, nous effectuons une validation croisée en utilisant 10 groupes de photos.

Il est important de créer ces échantillons aléatoirement, de façon à pouvoir répéter les expériences et observer la reproductibilité des résultats. En effet si les erreurs sont deux fois plus importantes d'un essai à l'autre, cela signifie que le modèle est très fortement sensible au choix des images et que la capacité de généralisation du modèle est faible. Pour tous les résultats présentés dans ce document, nous répétons la validation croisée 10 fois afin de présenter une moyenne des performances obtenues, et donc de s'assurer que les résultats sont reproductibles.

2.4.2.2 Évaluation de la classification

Nous avons besoin de définir des critères objectifs permettant de quantifier la performance des modèles créés. Dans le cas de la classification, le critère le plus simple et le plus intuitif correspond au taux de bonne classification. Nous notons ce taux T_{BC}, et celui-ci désigne le rapport entre le nombre d'images correctement classées et le nombre total d'images noté N_I.

Classification binaire

Dans le cas d'un problème à 2 classes, en plus du taux de bonne classification, il est courant d'utiliser la courbe ROC, pour *Receiver Operating Characteristic*. Cette courbe permet d'avoir une idée de la performance du classifieur lorsque le seuil de discrimination varie. Par exemple, dans le cas d'une classification obtenue par SVM, les images sont classées en fonction de leur position par rapport aux vecteurs de support. Plus la distance entre une image et ces vecteurs est élevée, plus les chances de l'image d'être correctement classée sont grandes. En augmentant progressivement la distance minimale nécessaire entre l'image et la marge permettant d'associer l'image à la classe, le taux de faux positifs (images évaluées à tort comme faisant partie de la classe) diminue et le taux de faux négatifs (images évaluées à tort comme ne faisant pas partie de la classe) augmente. Ce principe est illustré sur la figure 2.8. Un critère de pertinence du classifieur dans le cas de la classification binaire est ainsi défini par l'aire sous la courbe ROC, que nous notons *AUC* pour *Area Under the Curve*. A partir de cette courbe, d'autres critères peuvent être définis, comme la valeur T_{EQ} pour laquelle les taux de faux positifs et de faux négatifs sont égaux. Cette dernière valeur ne sera pas utilisée dans nos expériences, car elle est généralement très proche du taux de bonne classification T_{BC}. Les deux valeurs sont d'ailleurs égales lorsque le classifieur ne fait pas plus d'erreurs dans une classe que dans une autre.

Classification à plusieurs catégories

Lorsque plus de deux catégories sont considérées, il est non seulement intéressant d'étudier le taux de bonne classification, mais également l'importance de chaque erreur. En effet, il est plus grave de confondre une image dont l'évaluation est très mauvaise avec une image très bien évaluée qu'avec une image moyenne. Pour estimer le poids de chaque erreur, nous utilisons deux critères. Les mesures d'erreur entre catégories, notées $CCE(k)$ pour *Cross-Category Error*, permettent de quantifier le nombre d'images surévaluées de k catégories. $CCE(0)$ est

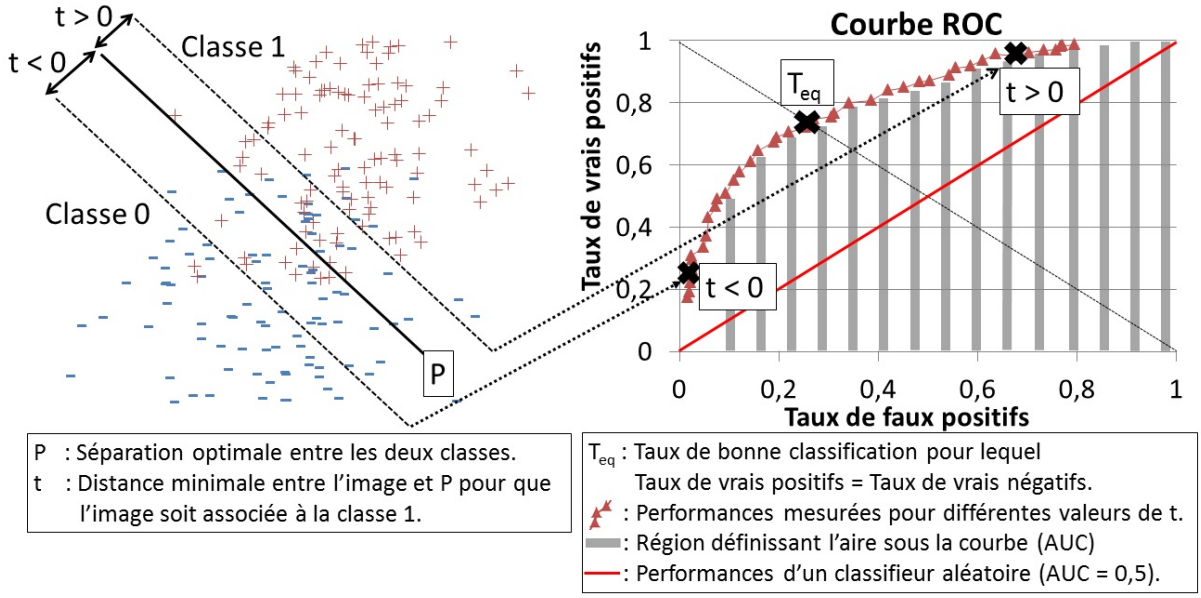


FIGURE 2.8 – Principe de construction d'une courbe ROC pour une classification par SVM. Chaque point de la courbe est obtenu à partir d'une valeur particulière du seuil t, représentant la distance nécessaire entre une image et l'hyperplan séparateur pour qu'une image soit associée à la classe 1.

ainsi le nombre d'images bien classées et $CCE(k)$ est le nombre d'images surévaluées de k catégories. Nous considérons aussi la somme des erreurs pondérées par leur importance et nous notons ce critère MCE pour *Multi-Category Error*. Les formules permettant de calculer ces valeurs sont indiquées dans les équations 2.15 et 2.16. Nous notons N_t le nombre d'images testées et C le nombre de classes.

$$\forall k \in [\![-(C-1),C-1]\!], CCE(k) = \sum_{i=1}^{N_t} \mathrm{Ind}(\hat{c}_i - c_i = k) \tag{2.15}$$

Dans l'équation 2.15, c_i est la classe de l'image i correspondant à la vérité terrain et $\hat{c}_i$ la prédiction de l'algorithme. k est la différence entre la vérité terrain et la classe prédite. La fonction $Ind(.)$ prend la valeur 1 lorsque la condition $\hat{c}_i - c_i = k$ est remplie, 0 sinon. $CCE(k)$ est donc le nombre de cas dont la différence entre la vérité terrain et la prédiction est égale à k. A partir de l'équation 2.15, nous définissons :

$$MCE = \frac{1}{MCE_{\text{Alea}}} \sum_{k=-(C-1)}^{C-1} |k| CCE(k) \tag{2.16}$$

Le terme MCE_{Alea} est ajouté ici afin d'obtenir une valeur de MCE entre 0 et 1. Ce terme est la valeur obtenue par le MCE dans le cas où la prédiction est faite par un classifieur aléatoire. Dans ce document, nous répartissons les images dans chaque catégorie de manière à obtenir un nombre d'images identique dans chaque catégorie : $\forall c \in [\![0,C-1]\!], N_c = N_t/C$, où N_c est

le nombre d'images dans la classe c. Dans ce cas, l'expression de MCE_{Alea} est :

$$MCE_{\text{Alea}} = \frac{N_t}{C}\frac{C^2-1}{3} = \mathcal{O}(N_t C) \tag{2.17}$$

Un bon classifieur présente un $T_{BC} = CCE(0)/N_t$ proche de 1, et une mesure d'erreur MCE proche de 0. A l'inverse un taux de bonne classification proche de $1/C$ et une valeur de MCE proche de 1 signifient que le classifieur ne fait pas mieux que le hasard.

Conclusion

Nous venons de définir différentes mesures de performances pour évaluer la classification. Certaines apportent des informations dans le cas de la classification binaire (courbe ROC), d'autres permettent de savoir précisément le nombre et le poids de chaque erreur dans le cas de la classification à plus de 2 classes (répartition des valeurs du CCE). Dans ce dernier cas, l'étude conjointe des mesures de T_{BC} et de MCE permet d'évaluer les performances des algorithmes selon des objectifs précis. Certains algorithmes fournissent en effet des résultats très précis et renvoient généralement un T_{BC} élevé, quitte à faire des erreurs plus importantes (MCE élevé), tandis que d'autres ont tendance à renvoyer des taux de bonne classification plus faibles mais font des erreurs de plus faible amplitude (MCE proche de 0). Une méthode permettant de tirer profit de ces informations est proposée en 2.5.2.

2.4.2.3 Régression

Les critères utilisés précédemment pour l'évaluation des résultats de classification ne sont pas pertinents pour la régression. Tout d'abord, l'erreur quadratique moyenne EQM est considérée et permet de quantifier l'écart entre les scores de vérité terrain de l'image i, s_i, et les prédictions, $\hat{s}_i$. Comme précédemment, N_t est le nombre d'images testées et nous pouvons calculer EQM à l'aide de la formule suivante :

$$EQM = \sqrt{\frac{1}{N_t}\sum_{i=1}^{N_t}(\hat{s}_i - s_i)^2} \tag{2.18}$$

Toutefois cet indice n'est pas suffisant pour estimer la qualité d'une régression. En effet, selon la répartition et la valeur des scores de vérité terrain, EQM peut prendre des valeurs très différentes. De plus, si les scores de vérité terrain sont très resserrés (écart-type faible), une prédiction constante aura tendance à fournir une EQM particulièrement faible, alors qu'un modèle cherchant à évaluer correctement les valeurs extrêmes aura tendance à faire des erreurs plus importantes. Pour remédier à ce problème, nous utilisons le coefficient de Pearson, noté R, défini par :

$$C_R = \frac{\sum_{n=1}^{N_t}(\hat{s}_i - \bar{\hat{s}}) \cdot (s_i - \bar{s})}{\sqrt{\sum_{n=1}^{N_t}(\hat{s}_i - \bar{\hat{s}})^2} \cdot \sqrt{\sum_{n=1}^{N_t}(s_i - \bar{s})^2}} \tag{2.19}$$

où $\bar{s} = \frac{1}{N_t}\sum_{n=1}^{N_t} s_i$ et $\bar{\hat{s}} = \frac{1}{N_t}\sum_{n=1}^{N_t} \hat{s}_i$ sont respectivement les moyennes de vérité terrain et de prédiction. Enfin, nous utilisons également la corrélation de Spearman ρ qui se calcule par

$$\rho = 1 - \frac{6\sum_{n=1}^{N_t} d_i^2}{N_t(N_t^2 - 1)} \tag{2.20}$$

où $d_i = rang(\hat{s}_i) - rang(s_i)$ est la différence de classement entre les variables. Le coefficient de Pearson mesure combien les variables sont linéairement corrélées, tandis que la corrélation de Spearman décrit simplement si les données sont corrélées de manière monotone. Dans les deux cas, des valeurs proches de 1 proviennent de données fortement corrélées tandis que des variables totalement décorrélées auront des coefficients proches de 0.

2.5 Post-traitement - Fusion des scores

Nous présentons dans cette partie une méthode permettant d'accroître la robustesse des prédictions effectuées à l'aide des 4 algorithmes d'apprentissage décrits dans les parties précédentes. L'idée de cette méthode est de fusionner les résultats proposés par les différents algorithmes d'apprentissage. Par exemple, dans le cas de la régression, plutôt que de considérer la prédiction d'un seul algorithme, nous pouvons faire la moyenne des différentes sorties de chaque algorithme. Cependant, différents biais sont introduits par les différences entre les modèles. Tout d'abord, comme nous l'avons déjà évoqué, certains algorithmes ont tendance à renvoyer des scores proches de la moyenne (forêts aléatoires, forêts boostées), tandis que d'autres ont tendance à présenter des écarts-types de scores plus élevés. Avant la fusion, les scores sont donc normalisés (voir 2.5.1) de manière à ce que les écarts-types des prédictions soient identiques pour chaque algorithme. En outre, selon les données, certains algorithmes peuvent être plus pertinents que d'autres. Pour tenir compte de cela, nous pondérons les poids de chaque prédiction par les performances de chaque algorithme, de manière à ce que les algorithmes aient moins d'influence sur les performances globales (voir 2.5.2).

2.5.1 Normalisation des scores pour la régression

Nous montrons sur la figure 2.9 différents nuages de points obtenus pour un même jeu de données en utilisant les 4 algorithmes définis en 2.4.1. Nous voyons sur ces images que non seulement les prédictions sont plus ou moins corrélées avec les scores de vérité terrain selon les algorithmes utilisés (différentes valeurs du coefficient de corrélation R), mais surtout que l'échelle des scores est plus ou moins utilisée selon les algorithmes. Si nous souhaitons fusionner les scores fournis par chaque algorithme afin d'effectuer une prédiction globale, il parait judicieux de procéder d'abord à une phase de normalisation de cette échelle des scores, de manière à ce que l'écart-type des prédictions soit le même pour chaque algorithme.

Nous notons respectivement M_{VT} et ET_{VT} la moyenne et l'écart-type des scores des images dans la base d'apprentissage, et M_A et ET_A la moyenne et l'écart-type des prédictions de

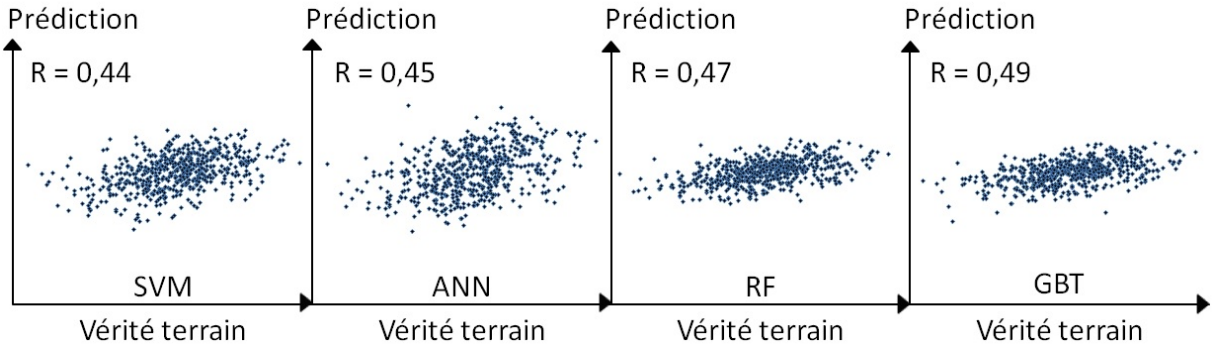


FIGURE 2.9 – Différents nuages de points obtenus à l'aide des 4 algorithmes présentés en 2.4.1. L'axe des abscisses représente les scores de vérité terrain, celui des ordonnées les scores prédits par chacun des algorithmes (respectivement pour chaque image : SVM, ANN, RF, GBT). Les scores de prédiction sont beaucoup moins dispersés pour RF et GBT.

l'algorithme A. Nous définissons pour chaque score s_i^A associé à l'image i et à l'algorithme A un score de prédiction ajusté $s_{i\text{Ajust}}^A$ à l'aide de la formule :

$$s_{i\text{Ajust}}^A = (S_i^A - M_A)\frac{\text{ET}_{VT}}{\text{ET}_A} + M_{VT} \tag{2.21}$$

Ainsi, les moyennes et les écarts-types des scores ajustés sont respectivement égaux à M_{VT} et ET_{VT} pour chaque algorithme, et nous obtenons les nuages de points renormalisés présentés sur la figure 2.10. Pour tous les résultats présentés dans ce document, nous utilisons les scores ajustés plutôt que les scores obtenus directement. Plusieurs remarques peuvent être faites concernant les changements induits par cette renormalisation :

- Dans la plupart des cas, la différence entre M_A et M_{VT} est négligeable : les algorithmes ont tendance à effectuer des prédictions dont la moyenne correspond à la moyenne des scores de vérité terrain.
- L'écart-type des scores de prédiction non normalisés est inférieur à l'écart-type des scores de vérité terrain. La normalisation proposée a donc tendance à accroître la dispersion des scores. De plus, les jeux de données proposés contiennent un grand nombre d'images moyennes, et peu d'images de score extrême. Ainsi, accroître la dispersion des scores a tendance à améliorer la prédiction des images extrêmes (qui nous intéressent particulièrement) et à diminuer la précision de la prédiction pour les images moyennes. Ces dernières étant plus nombreuses, la renormalisation augmente la moyenne des erreurs de prédiction et donc la valeur de l'erreur EQM.
- La renormalisation n'a aucune influence sur les coefficients de corrélation R et ρ.

2.5.2 Fusion des scores pour la classification et la régression

Afin d'obtenir une prédiction robuste, nous cherchons à combiner les prédictions de chaque algorithme. Le détail des combinaisons effectuées pour la classification et la régression est donné dans les paragraphes suivants.

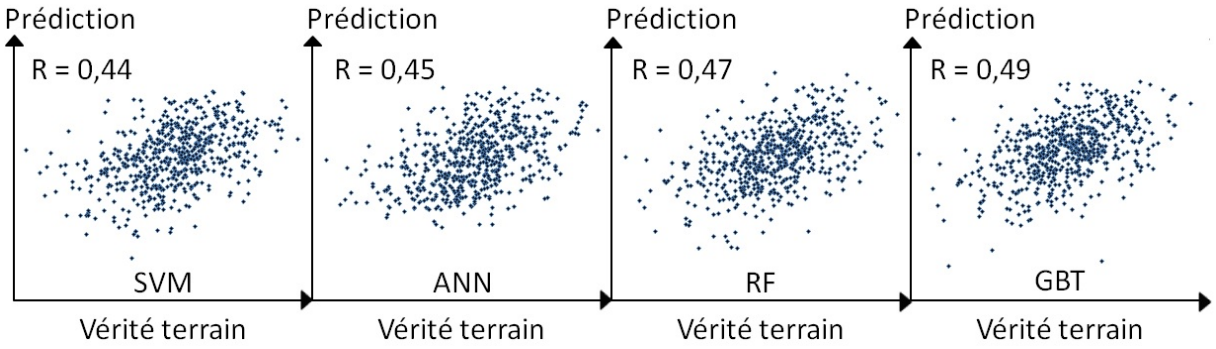


FIGURE 2.10 – Nuages de points obtenus à l'aide des 4 algorithmes présentés en 2.4.1, après normalisation des scores. Les scores de prédiction ont le même écart-type pour chaque algorithme. Des valeurs de prédiction plus éloignées de la moyenne apparaissent, et les sorties des algorithmes sont plus facilement comparables.

Classification

Soit $c_A(i)$ la classe prédite pour l'image i à l'aide de l'algorithme A, et $c(i)$ la prédiction obtenue par combinaison des différentes classes $c_A(i)$. Nous pouvons évaluer les performances de la classification de l'algorithme A à l'aide du taux de bonne classification $T_{BC}(A)$ et des erreurs pondérées par la différence entre les classes attendues et prédites, pour lequel nous avons défini le critère noté $MCE(A)$. Les valeurs $T_{BC}(A)$ et $MCE(A)$ sont calculées par validation croisée sur la base d'apprentissage. Nous fusionnons finalement les différentes prédictions $c_A(i)$ à l'aide de l'équation 2.22.

$$c(i) = \frac{\sum_A c_A(i)(\frac{T_{BC}(A)}{MCE(A)})^p}{\sum_A (\frac{T_{BC}(A)}{MCE(A)})^p} \tag{2.22}$$

Remarquons tout d'abord dans l'équation 2.22 l'utilisation du rapport $T_{BC}(A)/MCE(A)$. Si ce rapport est élevé, l'algorithme A est performant et son poids dans la prise de décision est important. Le paramètre p est choisi de manière à augmenter ou diminuer l'impact du poids de chaque algorithme. En effet, il arrive qu'un algorithme particulier soit significativement plus performant que les autres, et dans ce cas ajouter d'autres algorithmes dans la prise de décision risque de diminuer les performances globales. Nous avons constaté que pour améliorer la performance globale, les différents algorithmes pris en compte doivent présenter des performances proches, de l'ordre de 10% d'écart relatif entre leurs rapports $T_{BC}(A)/MCE(A)$. Pour cela, nous avons choisi de fixer la valeur de p à 10. Ainsi, si deux algorithmes A et B présentent des performances telles que $(T_{BC}(A)/MCE(A))/(T_{BC}(B)/MCE(B)) = 0{,}9$, le poids de l'algorithme B dans la prise de décision sera 3 fois supérieur à celui de A. Lorsque l'écart relatif est de l'ordre de 20%, le ratio entre les deux poids est de 10 : l'algorithme le moins performant n'est alors quasiment pas pris en compte.

Régression

Le principe de la fusion de scores dans le cas de la régression est le même. Notons $s_A(i)$ le score de prédiction pour l'image i à l'aide de l'algorithme A, et $s(i)$ la prédiction obtenue par combinaison des différents scores $s_A(i)$. Plutôt que le ratio $T_{BC}(A)/MCE(A)$, nous utilisons ici le coefficient de corrélation $R(A)$ pour pondérer les différents scores. Nous conservons le paramètre p défini précédemment et obtenons ainsi une prédiction globale dont l'expression est donnée dans l'équation 2.23.

$$s(i) = \frac{\sum_A s_A(i) R^p(A)}{\sum_A R^p(A)} \tag{2.23}$$

Nous ne considérons dans nos travaux que 4 algorithmes, mais les formules définies dans les équations 2.22 et 2.23 sont tout à fait transposables au cas où plus d'algorithmes seraient utilisés. Il est également envisageable d'utiliser plusieurs fois le même algorithme avec différents paramètres puis de fusionner les scores produits par chaque variante afin d'améliorer la robustesse des résultats. Nous présentons l'augmentation des performances obtenues par la fusion de ces scores dans les prochains chapitres de ce document, dans lesquels nous faisons référence à cette technique par l'abréviation LSF, pour "Late Score Fusion". En appliquant cet algorithme au jeu de données présenté sur les figures 2.10 et 2.9, nous obtenons le nuage de points présenté sur la figure 2.11. La corrélation entre les scores de vérité terrain et de prédiction est supérieure à celle obtenue en n'utilisant qu'un seul algorithme ($R = 0{,}51$ contre $R = 0{,}49$ pour le meilleur des algorithmes pris séparément). Ce résultat confirme la pertinence de la fusion proposée.

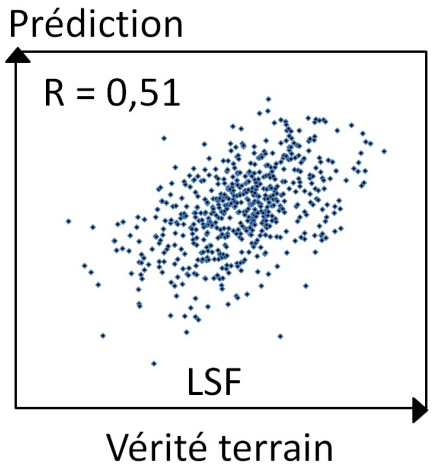


FIGURE 2.11 – Nuage de points obtenu par de la fusion des 4 algorithmes présentés en 2.4.1 (LFS). L'axe des abscisses représente les scores de vérité terrain, celui des ordonnées les scores prédits par la fusion des algorithmes.

2.6 Conclusion

Nous venons de définir les différentes étapes que nous considérons afin de construire un modèle d'évaluation des images :

1. La représentation et l'analyse des données,
2. L'utilisation d'algorithmes d'apprentissage automatique,
3. L'évaluation des performances de prédiction.

Différentes améliorations par rapport aux schémas classiques d'apprentissage supervisé et de prédiction ont été proposées. Tout d'abord, nous avons adapté l'algorithme Relief à nos jeux de données afin d'obtenir un retour sur la pertinence de chacune des caractéristiques que nous extrayons. Nous montrons dans le chapitre 3 que cet algorithme permet également d'améliorer les performances de la prédiction en supprimant les informations peu utiles des vecteurs de caractéristiques. Après l'apprentissage, nous procédons à la fusion des prédictions de chaque algorithme en combinant leur évaluation. Cette combinaison des prédictions permet d'améliorer les performances de prédiction.

Le cadre de l'analyse de données et de l'apprentissage statistique proposé dans ce chapitre a été conçu afin d'améliorer les performances de prédiction dans le cas de données fortement bruitées (évaluations subjectives), complexes (nombreux critères à prendre en compte) et réduites (seulement quelques centaines à quelques milliers d'images). De nombreux paramètres ont dû être fixés, et souvent ces choix ont été fait empiriquement (de meilleurs résultats ont été observés). Le cadre proposé est suffisamment générique pour qu'il puisse être adapté à tous les jeux de données et de caractéristiques dont nous allons discuter dans ce document. C'est pour cette raison que nous n'avons pas détaillé la phase d'extraction de caractéristiques. Cette étape est en effet dépendante de l'objectif et sera décrite dans les deux chapitres suivants, traitant respectivement de l'estimation de la qualité esthétique et des impressions de compétence et de sympathie dégagées par une photo de visage. Les paramètres définis dans ce chapitre sont adaptés à nos données, mais il n'est pas garanti que leur choix soit optimal pour n'importe quel jeu de données. Nous donnons quelques exemples de situations où les paramètres définis dans ce chapitre ont dû être modifiés afin d'obtenir de meilleures performances.

Chapitre 3

Estimation de la qualité esthétique d'une photo de visage

Sommaire

3.1 Introduction . . . 78
3.2 Bases d'images annotées . . . 79
3.2.1 Human Face Scores (HFS) . . . 80
3.2.2 Photo Net Faces (PNF) . . . 81
3.2.3 Face Aesthetic Visual Analysis (FAVA) . . . 83
3.2.4 Photo Quality (CUHK - PQ) . . . 84
3.2.5 Flickr . . . 85
3.2.6 Bilan : Résumé des bases considérées . . . 86
3.3 Définition des descripteurs considérés . . . 87
3.3.1 Attributs de bas niveau . . . 88
3.3.2 Composition de l'image . . . 96
3.4 Définition des régions du visage considérées . . . 97
3.4.1 Extraction des différentes zones du visage . . . 97
3.4.2 Performances de la détection des attributs . . . 101
3.4.3 De la détection à la définition des régions considérées . . . 103
3.5 Étude de la pertinence des descripteurs et des régions . . . 104
3.5.1 Influence des caractéristiques . . . 105
3.5.2 Influence des régions . . . 107
3.5.3 Bilan : caractéristiques et régions pertinentes . . . 107
3.6 Étude de la pertinence des algorithmes d'apprentissage supervisé . . . 109
3.7 Estimation de la qualité esthétique et validation du modèle . . . 111
3.7.1 Classification binaire . . . 111
3.7.2 Classification étendue à 3 catégories . . . 113
3.7.3 Régression . . . 114
3.8 Comparaison avec l'état de l'art . . . 115
3.8.1 Performances de classification . . . 115
3.8.2 Performances de régression . . . 120
3.9 Application à la recherche et la sélection de photographies de visage 121
3.9.1 Protocole proposé . . . 121
3.9.2 Photos sélectionnées . . . 122
3.9.3 Temps de calcul . . . 122
3.10 Conclusion . . . 125

3.10.1 Avantages de la méthode 125
3.10.2 Limites de la méthode 126
3.10.3 Bilan 126

3.1 Introduction

Dans ce chapitre, nous utilisons les outils introduits dans le chapitre 2 afin d'évaluer automatiquement la qualité esthétique de photos de visage. Nous détaillons en particulier les caractéristiques que nous extrayons sur les images avant de procéder à l'apprentissage statistique, et nous montrons qu'il est possible d'obtenir une évaluation pertinente de la qualité esthétique des photos à partir de ces caractéristiques. Différentes applications découlent des méthodes proposées : classement d'images par ordre de qualité esthétique (moteurs de recherche, tri de photos privées), aide à la capture d'images (intégration dans un appareil photo ou une webcam) ou édition automatique de photographies.

Pour réaliser cet objectif, nous nous restreignons à l'utilisation de descripteurs décrivant chacun un aspect particulier de la qualité esthétique de l'image. La photo est-elle nette ? Bien éclairée, contrastée ? Le choix des couleurs est-il adapté ? De nombreuses caractéristiques décrites dans les travaux précédents sur la qualité esthétique des photos ont été testées, et nous présentons dans ce chapitre celles qui nous permettent d'obtenir les meilleures performances de classification et de régression.

Nous proposons une méthode adaptée aux photos de visage. Cette méthode consiste à calculer chaque descripteur sur plusieurs zones de l'image, correspondant à l'image entière, au visage, ainsi qu'aux régions contenant les yeux et la bouche. Ce calcul de descripteurs à différentes échelles permet d'inclure implicitement différentes informations relatives à la composition de l'image : mise en valeur du visage, différences entre avant-plan et arrière-plan, etc.

Nous avons choisi de ne pas utiliser de critères de plus haut niveau tels que la présence de sourire ou l'analyse des expressions faciales car nous pensons que ces éléments sont dépendants du contexte dans lequel la photo est étudiée : un sourire est attendu pour une photo de vacances, un air sérieux est nécessaire pour une photo de CV. Les critères propres à une personne (estimation de l'âge, du sexe, de la couleur de peau, beauté du visage) ne sont pas considérés car nous cherchons à estimer la qualité esthétique d'une photo indépendamment de la personne. Nous n'utilisons pas les descripteurs d'image génériques tels que ceux proposés par [Marchesotti et Perronnin 2011] car nous nous limitons à l'utilisation de caractéristiques interprétables visuellement.

Le plan de ce chapitre est le suivant. Tout d'abord, nous décrivons en 3.2 les différentes bases de photographies que nous utilisons dans nos travaux. En particulier, nous avons construit 2 bases d'images annotées ; la première est issue d'une expérience en laboratoire,

la seconde est obtenue en récupérant un grand nombre d'images sur un site de partage de photographies. Puis, nous définissons en 3.3 les différentes caractéristiques que nous extrayons sur les photos. Celles-ci sont construites de manière à pouvoir être calculées sur les différentes régions proposées en 3.4. Dans chacune de ces régions, il devient alors possible de calculer des descripteurs décrivant différents aspects de l'image : illumination du visage, textures de l'arrière-plan, etc. Une analyse des caractéristiques et des régions pertinentes est proposée en 3.5. Nous étudions l'influence des différents algorithmes d'apprentissage en 3.6. À l'aide de ces informations, nous estimons enfin la qualité esthétique de photographies en 3.7 puis nous comparons nos travaux à l'état de l'art en 3.8. Enfin, nous concluons ce chapitre par une application de notre travail à la recherche et la sélection de photographies de visage en 3.9.

Nos principales contributions à l'estimation automatique de la qualité esthétique de photos de visage sont les suivantes :

— Création de deux bases de photos, pour lesquelles les images sont évaluées selon leur qualité esthétique. La première base (250 images) contient des photos évaluées dans un environnement contrôlé, tandis que la seconde comprend plus de 80000 photos contenant des visages et récupérées automatiquement sur le site [*PhotoNet*].
— Définition de caractéristiques décrivant des informations globales (par exemple sur l'éclairage, la netteté) sur une région particulière de la photo.
— Définition de régions de calcul des caractéristiques adaptées à l'évaluation de photos de visage. En effet, les travaux de l'état de l'art se limitent généralement à l'extraction de caractéristiques sur l'image entière et sur le visage entier. Nous proposons d'étudier des régions plus précises, telles que celles définies par les yeux ou la bouche. En particulier, l'extraction de caractéristiques uniquement sur la région des yeux, très riche en information, permet d'obtenir de très bonnes performances alors que la région est très petite par rapport à la taille de l'image. Ce résultat permet d'accélérer significativement l'évaluation dans le cadre d'une utilisation en temps réel.
— Utilisation des méthodes présentées dans le chapitre précédent (sélection de caractéristiques, fusion des prédictions de chaque algorithme) afin d'améliorer la précision de nos estimations.

3.2 Bases d'images annotées

Différentes catégories de bases de photos de visage annotées peuvent être distinguées. Ces différences peuvent porter sur :

— La méthode d'évaluation des images. Par exemple, des images évaluées par des humains en laboratoire selon des critères clairement énoncés, avec un écran dont les paramètres d'illumination et de couleur sont figés, ainsi qu'une distance de visionnage identique pour tous, améliorent la fiabilité des scores obtenus. A l'inverse, des images récupérées automatiquement sur des sites de partage de photos ([*PhotoNet*], [*DPChallenge*]) et notées par des internautes, présentent plus de biais de notation : nombre de votes très divers, seule une certaine catégorie de photos est représentée (généralement les photos les plus réussies) et seuls les utilisateurs du site évaluent les photos.

— Le nombre d'images dans la base. Il est très difficile de faire évaluer un grand nombre d'images par un grand nombre de personnes dans un environnement très contrôlé (cela coûte du temps et de l'argent). Généralement, ce type de bases d'images ne dépasse pas quelques centaines de photos.
— Les contraintes sur les images. Nous nous intéressons essentiellement aux photos de visage avec les contraintes définies en introduction de ce document : visages centrés, de face, suffisamment grands. Toutefois, la plupart des travaux antérieurs [Li et al. 2010a ; Khan et Vogel 2012 ; Redi et al. 2015] ne posent pas de contraintes aussi strictes, et considèrent toutes les images contenant au moins un visage.
— La moyenne de qualité esthétique globale des images. Typiquement, une base constituée de photos privées prises par des photographes amateurs présente des images de plus faible qualité esthétique qu'une base créée par des photographes professionnels.

Lorsque nous décrivons les différentes bases d'images que nous considérons, nous indiquons comment les images se situent par rapport aux critères définis ci-dessus. Nous donnons en outre des exemples d'images, ainsi que des informations sur la présence d'images en noir et blanc, la taille moyenne des images (ce paramètre peut fortement varier d'une base à l'autre), l'origine des images, et la distribution des scores des images.

En plus de celles présentes dans l'état de l'art, nous décrivons ici deux bases que nous avons construites afin de valider les méthodes décrites dans ce document. La première, que nous nommons HFS (pour "Human Face Scores"), est une base de 250 images évaluées dans un environnement contrôlé. La seconde a été obtenue à partir du site [*PhotoNet*], à partir duquel nous avons récupéré les images évaluées par au moins 5 personnes et contenant des visages. Nous obtenons au total environ 80000 images, dont plus de 28000 correspondent à nos contraintes (visages centrés suffisamment grands). Nous notons cette seconde base PNF, pour "Photo Net Faces". Nous utilisons également d'autres bases dans nos travaux, essentiellement pour comparer nos travaux avec l'état de l'art et afin de nous assurer que les modèles que nous proposons fonctionnent pour différents types de photographies.

3.2.1 Human Face Scores (HFS)

Origine des photos

Au total, la base HFS contient 250 photos, parmi lesquelles figurent deux sous-ensembles. Ainsi, 20 personnes sont représentées chacune exactement 7 fois, ce qui constitue un premier ensemble de 140 photos. L'objectif final, discuté dans les chapitres 4 et 5, est de tester la capacité du programme à déterminer quelle sera la photo la plus adaptée pour un individu et une application donnés. Les 110 photos restantes sont issues de personnes différentes, provenant de sources diverses.

Une bonne partie des photos proviennent de la base de données disponible sur `http://www.vision.caltech.edu/html-files/archive.html`. Cette base propose des photos de visage de face, et chaque individu est représenté plusieurs fois. Les éclairages et décors varient, ces photos constituent ainsi une bonne partie du lot de 140 photos décrivant plusieurs fois une

même personne. Une autre source de photos de visage est donnée par la base LFW disponible sur `http://vis-www.cs.umass.edu/lfw/` [Huang et al. 2007], que nous avons déjà présentée en 3.4.2. Enfin, quelques photos sont extraites de collections privées.

Évaluation des photos

Les 250 photos sont évaluées par un ensemble de 25 observateurs, majoritairement français et âgés de 20 à 55 ans. L'expérience est effectuée dans la même pièce et dans les mêmes conditions pour tous les participants : la distance à l'écran et sa luminosité sont fixées. Les participants ne sont pas familiers avec l'objectif de l'expérience, et ont reçu pour consigne : *« Vous allez voir des photos de visage apparaître à l'écran. Pour chaque photo, merci de juger la QUALITÉ ESTHÉTIQUE GÉNÉRALE DE LA PHOTO. Une photographie sera considérée comme très esthétique lorsque les nombreux aspects qui la caractérisent (cadrage, luminosité, résolution, contraste, flou, équilibre entre les éléments de la composition, etc.) sont de bonne qualité. Merci d'indiquer à l'aide du clavier sur une échelle allant de 1 à 6 la QUALITÉ ESTHÉTIQUE DE LA PHOTO avec 1 : très mauvaise qualité et 6 : excellente qualité. »*

Les participants doivent s'entraîner sur un ensemble de photos ne faisant pas partie de la base avant le début de l'expérience, afin d'appréhender l'échelle des notes et de se faire une idée du type de photographies qu'ils ont à évaluer. Après l'expérience, les notes sont moyennées sur l'ensemble des votants pour chaque photo. Chaque photo est alors associée à un score correspondant à la note moyenne des 25 participants. Ce score est défini comme étant la vérité terrain, et est l'objectif que nous cherchons à atteindre à l'aide de nos modèles de prédiction.

La répartition des notes est donnée sur la figure 3.2. La moyenne des scores sur l'ensemble des photos est de 3,21 et l'écart type de 0,73. L'échelle des scores est entièrement utilisée (les scores vont de 1,36 à 5,36). Des exemples d'images sont donnés sur la figure 3.1. Finalement, les images présentées ici sont très strictes dans leur composition (hauteur normalisée à 240 pixels, visage centré et de même taille), la base contient très peu d'images, les photos sont plutôt des photos d'amateurs de moyenne qualité esthétique. Il n'y a pas d'images en noir et blanc dans cette base.

3.2.2 Photo Net Faces (PNF)

HFS présente l'inconvénient de ne réunir que 250 images. Or, il est difficile de proposer un modèle performant et généralisable à partir de si peu d'images, et nous avons pour cela créé une autre base de photos, plus conséquente, à partir du site [*PhotoNet*]. Nous avons choisi ce site car à notre connaissance, il n'existe pas de base à très grande échelle (plusieurs centaines de milliers de photos) reposant sur les images de [*PhotoNet*]. Pour construire cette base, nous avons procédé selon le protocole suivant.

1. Nous commençons par analyser automatiquement toutes les pages du site [*PhotoNet*],

FIGURE 3.1 – Exemples d'images présentes dans la base HFS.

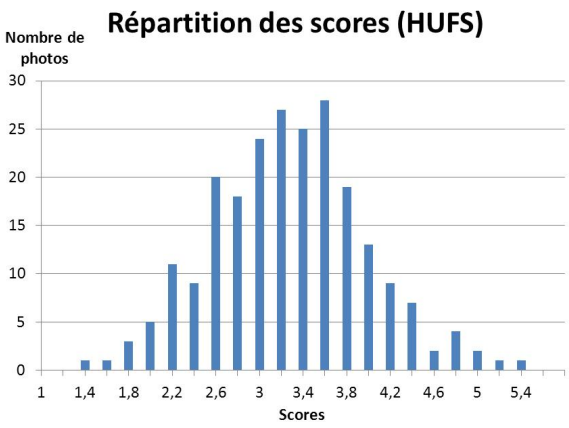


FIGURE 3.2 – Répartition des scores pour la base HFS.

en ne considérant que celles contenant des images évaluées par au moins 5 personnes. Nous avons recensé plus d'un million d'images correspondant à ces critères, que nous avons téléchargées et pour lesquelles nous récupérons également le nombre de votes ainsi que la note moyenne.

2. Nous utilisons l'algorithme de détection de visage proposé par [Viola et Jones 2001] afin de ne conserver que les photos contenant des visages. Nous obtenons alors environ 100000 images, parmi lesquelles figurent de nombreuses fausses détections ; nous avons en effet utilisé des critères de détection assez souples de manière à obtenir un nombre maximal d'images contenant des visages.

3. Nous vérifions alors manuellement que les images obtenues lors de l'étape précédente ne sont pas des faux positifs. Après vérification et suppression des faux positifs, nous conservons environ 82000 photos contenant des visages.

4. Nous trions enfin les photos de visage de façon à nous limiter à celles respectant les contraintes de composition établies précédemment. Finalement, nous disposons de 28272 photos de visage respectant nos contraintes, et nous notons cette base PNF, pour Photo Net Faces. Certaines de ces images sont en noir et blanc.

Dans ce document, nous travaillons essentiellement à partir des 28000 images respectant nos contraintes. Nous présentons tout de même une application possible de notre travail à la sélection automatique de photos de visage en exploitant la base de 82000 photos dans le chapitre final. Si la base que nous venons de créer présente l'avantage déterminant de contenir un très grand nombre de photos, certaines limites de cet ensemble sont à souligner.

La très grande majorité des images est évaluée par un faible nombre de personnes : plus de 90% des images sont évaluées par moins de 10 personnes. Il convient alors de se poser la question de la pertinence du score moyen résultant de ces votes. Il est par exemple intéressant de noter qu'il existe une corrélation importante entre le nombre de votes correspondant à une image et sa moyenne de qualité esthétique : il semble naturel que les images les plus esthétiques soient vues et partagées par un plus grand nombre de personnes.

Aussi, PNF contient essentiellement des photos d'excellente qualité esthétique provenant de photographes parfois professionnels, et la distribution des scores est très resserrée (écart-type de 0,63, voir figure 3.4) autour d'une note moyenne déjà très élevée (4,81 sur 7). Ceci rend très difficile la distinction entre deux catégories d'images associées respectivement aux notes au-dessus et en-dessous du score médian (4,83). Cette base est ainsi plus adaptée à des modèles de classification dont l'objectif est de distinguer des images de très bonne ou de très mauvaise qualité (voir CUHKPQ en 3.2.4), ou encore aux problèmes de régression dont l'objectif est d'estimer le score de vérité terrain associé aux images.

FIGURE 3.3 – Exemples d'images présentes dans la base PNF.

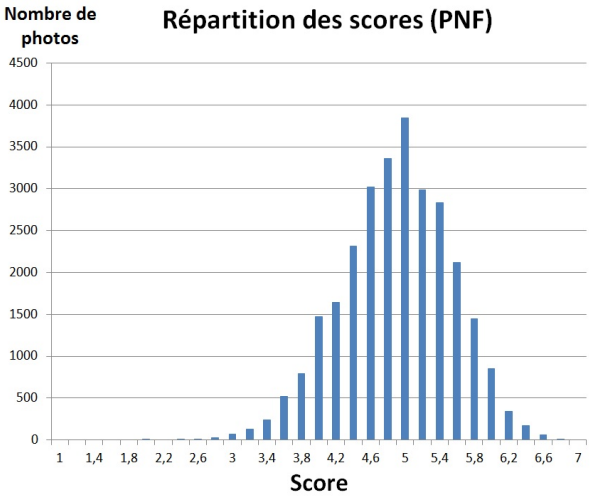


FIGURE 3.4 – Répartition des scores pour la base PNF.

3.2.3 Face Aesthetic Visual Analysis (FAVA)

Origine et évaluation des photos

La base AVA, décrite dans [Murray et al. 2012], contient des photos récupérées automatiquement sur le site `http://www.dpchallenge.com` [*DPChallenge*]. Celles-ci proviennent de photographes amateurs ou professionnels, et sont évaluées par les utilisateurs du site selon leur qualité esthétique. Les photos sont classées par "challenge", c'est à dire que les photographes participent à des concours de photos correspondant à des thèmes particuliers. Ensuite, un système de vote permet à chacun de donner un score aux photos, sur une échelle de 1 à 10, 10 étant la note maximale.

Création de la sous-base FAVA

La base AVA contient au total plus de 250000 images correspondant à un millier de "challenges". Pour constituer la base FAVA, nous avons parcouru automatiquement l'ensemble des 250000 images et conservé celles contenant des visages correspondant à nos contraintes (vi-

sages suffisamment grands, centrés et de face, photos en couleurs). Seules 636 photos ont ainsi été retenues, pour un total d'environ 10000 photos contenant des visages.

Celles-ci ont toutes été évaluées par au moins 78 utilisateurs de [*DPChallenge*], ce qui assure une certaine fiabilité des scores moyens, qui définissent la vérité terrain pour chaque image. La répartition des scores de FAVA est donnée sur la figure 3.6. Nous voyons que l'échelle des scores n'est pas entièrement utilisée, les notes allant de 2,8 à 7,6. L'ensemble des photos possède une moyenne de 5.4 pour un écart-type de 0.65. Comparativement à HFS, l'écart type des scores moyens de la base FAVA est faible. La séparation automatique entre les photos les mieux notées et celles aux notes les plus faibles sera donc probablement plus difficile. Nous constatons également que les photos présentent le même biais que pour PNF : les images de mauvaise ou de très mauvaise qualité esthétique sont très rares, et le problème consistant à distinguer deux catégories d'images parmi ces photos revient à distinguer des photos correctes d'amateurs de photos réussies de professionnels. Des exemples d'images sont donnés sur la figure 3.5.

FIGURE 3.5 – Exemples d'images présentes dans la base HFS.

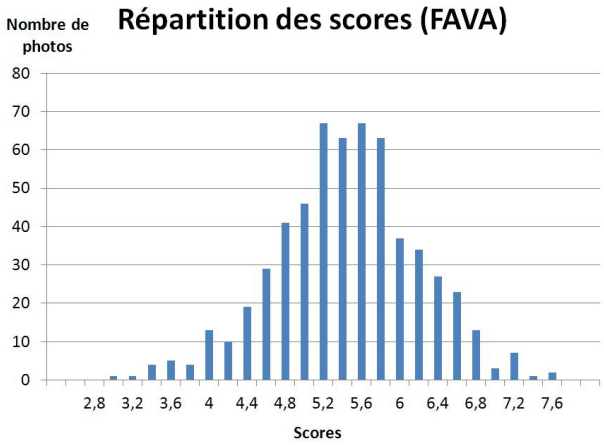


FIGURE 3.6 – Répartition des scores pour la base FAVA.

3.2.4 Photo Quality (CUHK - PQ)

La base CUHKPQ, pour "Chinese University of Hong Kong : Photo Quality", est présentée par [Tang et al. 2013] et permet d'établir des modèles de qualité esthétique pour différentes catégories d'images : paysages, portraits, animaux, architecture, etc. Nous ne nous intéressons cependant qu'à la catégorie des portraits, pour laquelle nous conservons les images dans lesquelles nous détectons au moins un visage.

La base CUHKPQ est construite en supprimant toutes les photos dont les scores de vérité terrain sont "moyens". Plus précisément, seules les photos pour lesquelles au moins 8 observateurs sur 10 estiment que l'image est de très bonne ou de très mauvaise qualité esthétique sont conservées. Les images sont ainsi classées en deux catégories correspondant respectivement aux images dont la qualité esthétique est soit très faible, soit très élevée. Cette base est

utilisée dans nos travaux essentiellement à des fins de validation de la méthode : les images étant très différentes entre les catégories, le problème de classification est plus facile. Cette base nous permet également de comparer nos résultats à d'autres méthodes de l'état de l'art. Des exemples d'images pour chaque catégorie sont proposés en figure 3.7. La distribution des scores n'est pas donnée car les images sont ici uniquement associées à une étiquette de classe (bonne ou mauvaise qualité esthétique). Enfin, notons que la composition des images est libre ; les visages ne respectent pas les contraintes de taille et de position définies en introduction.

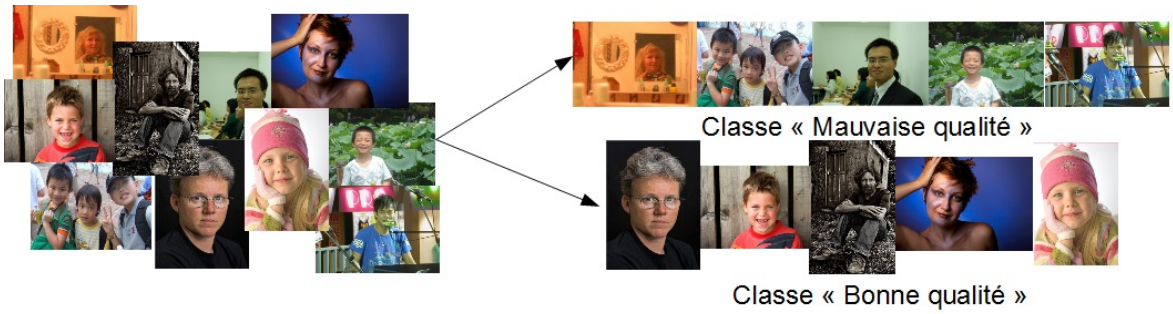


FIGURE 3.7 – Exemples d'images présentes dans la base CUHK. Les images sont soit des images de très bonne, soit de très mauvaise qualité esthétique.

3.2.5 Flickr

Une dernière base d'images est considérée, essentiellement dans l'objectif de situer les performances de nos algorithmes par rapport à d'autres travaux sur l'estimation de la qualité esthétique de photographies contenant des visages. Cette base est constituée à partir d'un échantillon de 500 images récupérées sur le site [*Flickr*]. Ces 500 photos contiennent toutes des visages, certaines sont des photos de groupe (amis, famille), d'autres des portraits. Afin d'associer ces photos à un score de qualité esthétique, [Li et al. 2010a] utilisent le service web proposé par Amazon [1]. Ce service permet de connecter des entreprises ou des laboratoires ayant besoin d'acquérir des données nécessitant de l'intelligence humaine (typiquement l'annotation d'images ou de vidéos, comme c'est le cas ici) à des particuliers qui sont rémunérés pour le travail effectué. Les images utilisées par [Li et al. 2010a] sont notées de 1 à 10 par plus de 40 personnes [2]. La moyenne des scores se situe à 5,9 et l'écart-type est de 1,7. Des exemples d'images ainsi que la distribution des scores pour cette base sont respectivement donnés sur les figures 3.8 et 3.9.

L'idée présentée dans l'article de Li et al. est d'évaluer les images en tenant compte également des relations entre les visages (émotions, proximité). Nous utilisons ces photos dans nos travaux afin de comparer les performances de nos caractéristiques par rapport à celles proposées par d'autres [Li et al. 2010a ; Xue et al. 2013]. Notre objectif est de montrer que nos caractéristiques suffisent à obtenir des performances satisfaisantes sur des photos de visage plus complexes, dans lesquelles nous extrayons uniquement les informations décrites en 3.3.

1. Voir `http://aws.amazon.com/fr/mturk/`.
2. Photos disponibles en ligne à l'adresse `http://chenlab.ece.cornell.edu/downloads.html`.

FIGURE 3.8 – Exemples d'images présentes dans la base Flickr.

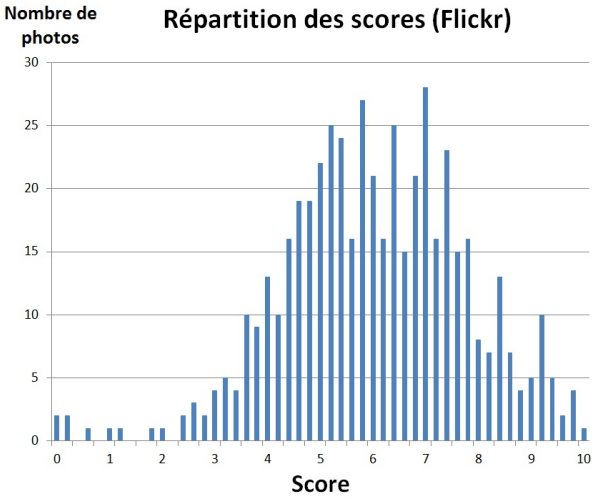


FIGURE 3.9 – Répartition des scores pour la base Flickr.

A partir de ces 500 photos, il est possible d'extraire le sous-ensemble d'images ne contenant qu'une seule personne. Ce sous-ensemble contient 145 photos et est utilisé par [Khan et Vogel 2012 ; Redi et al. 2015], dont les travaux portent sur l'étude des portraits et des méthodes permettant d'en évaluer la qualité esthétique. Nous comparons également les performances de nos caractéristiques avec ces travaux, et notons ce sous-ensemble FlickrP (pour Flickr Portraits).

3.2.6 Bilan : Résumé des bases considérées

A cause du biais existant entre le monde des photographes amateurs (photos de familles en vacances, photos prises avec un téléphone) et celui des photos partagées sur des réseaux sociaux pour photographes, il est difficile d'utiliser un modèle de qualité esthétique appris à partir d'une base d'images particulière pour prédire la qualité d'images d'une autre base. Par exemple, la construction d'un modèle à partir de PNF ne permet pas de prédire précisément la qualité esthétique des images de HFS. Nous pouvons constater que les bases de photos que nous définissons ici sont très différentes entre elles et répondent à des problématiques spécifiques. Dans le tableau 3.1, nous résumons les différentes propriétés des bases que nous venons de décrire.

Dans ce tableau, nous indiquons des informations sur les difficultés que nous avons rencontrées lors de la création de nos modèles de classification ou de régression. Typiquement, évaluer automatiquement des photos récupérées et évaluées dans des environnements très peu contrôlés (PNF, AVA) s'avère être une tâche très difficile. A l'inverse, lorsque les scores associés aux images sont très fiables (HFS) et que l'écart-type des scores est suffisamment élevé (CUHKPQ ne contient que des images d'excellente ou de très mauvaise qualité esthétique), il est possible d'obtenir des performances de prédiction plus élevées. La difficulté de chaque

Tableau 3.1 – Résumé des différentes bases de photos contenant des visages. Plus le nombre de + dans une case est élevé, plus les images de la base sont difficiles à évaluer (Difficulté) et les contraintes sur leur composition (Composition) sont élevées.

Nom de la base	Origine des images	Nombre d'images	Difficulté	Composition	Présentée dans
HFS	Diverse	250	+	++++	Ce travail
PNF	[*PhotoNet*]	28272	+++	+++	Ce travail
AVA	[*DPChallenge*]	10141	+++	-	Redi et al.
FAVA	[*DPChallenge*]	636	++	+++	Ce travail
CUHKPQ	[*DPChallenge*]	658	-	+	Tang et al.
Flickr	[*Flickr*]	500	++	-	Li et al.
FlickrP	[*Flickr*]	145	++	++	Khan et al.

base est indiquée par le nombre de + dans la colonne correspondante. Ce nombre est donné à titre indicatif, et est obtenu en comparant les performances de nos modèles sur chacune de ces bases : plus les performances sont faibles, plus la base est considérée comme difficile. Il est possible que d'autres méthodes d'estimation de la qualité esthétique [Li et al. 2010a ; Redi et al. 2015] rencontrent d'autres difficultés.

La colonne composition indique les contraintes sur la composition des images de chaque base. Par exemple, les bases pour lesquelles il est indiqué +++ sont celles pour lesquelles les photos respectent les contraintes définies en introduction de ce document. HFS présente une composition encore plus stricte car la taille des images est normalisée à 240 pixels de hauteur. Les bases associées à - sont celles pour lesquelles au moins un visage est visible, quelles que soient sa taille, sa position, son orientation, ou encore sa nature (visage de statue, de poupée...).

3.3 Définition des descripteurs considérés

Nous allons commencer à détailler les différentes caractéristiques que nous extrayons sur les images. Dans ce document, nous ne présentons que celles qui sont retenues dans le modèle de qualité esthétique final. Ces caractéristiques ont été sélectionnées en fonction de différents critères, et doivent ainsi :

1. Représenter une estimation d'un des aspects définissant la qualité esthétique de la photographie. Nous extrayons ainsi des indices de texture, de luminosité, de contraste, de couleur ou encore de composition de l'image. Nous souhaitons que les caractéristiques extraites puissent être interprétées directement par un utilisateur quelconque. Cela exclut par exemple les indices obtenus par l'utilisation de dictionnaires de mots visuels, qui traduisent la fréquence d'apparition de certains motifs sur l'image, et pour lesquels il n'est pas évident d'obtenir une interprétation. Cette contrainte permet d'obtenir automatiquement un retour sur les éléments qui influent sur la bonne ou la mauvaise qualité esthétique d'une image.

2. Ne pas tenir compte des conditions de visualisation de l'image, ou de ses métadonnées, auxquelles nous n'avons pas accès. De même, nous ne cherchons pas à extraire d'informations sur la personne prise en photo, et n'utilisons ainsi pas de données biométriques afin d'estimer la qualité esthétique de l'image.

3. Pouvoir être construites de manière à pouvoir être calculées sur n'importe quelle région de l'image. Cela nous permet d'introduire des informations à la fois locales et globales dans nos modèles, et surtout de pouvoir adapter les régions à des segmentations d'image particulières, propres aux photos de visage.

4. Permettre l'amélioration des performances globales des estimations. Il est toutefois possible que certaines caractéristiques soient pertinentes pour certaines bases d'images et moins pour d'autres, et c'est justement pour cette raison qu'il est intéressant d'utiliser un algorithme de sélection de caractéristiques.

Afin d'estimer la qualité esthétique de photographies de visage, nous proposons des attributs de bas niveau que nous regroupons en 5 catégories, correspondant à des indices de netteté, de texture, d'illumination, de contraste, et de couleur. Ces catégories ont été choisies car elles répondent aux différentes questions que peuvent se poser des observateurs humains : la photo est-elle nette ? Bien éclairée, contrastée ? Le choix des couleurs et des textures est-il pertinent ?

La majorité des travaux précédents [Li et al. 2010a ; Redi et al. 2015] considèrent des photographies ne respectant pas nos contraintes (visages de toutes tailles, non centrés). Ainsi, dans le but de comparer nos méthodes aux résultats de l'état de l'art, nous ajoutons à notre jeu de descripteurs des informations sur la composition de l'image.

3.3.1 Attributs de bas niveau

Les caractéristiques présentées dans cette partie peuvent être calculées sur une région quelconque de l'image. Toutefois, pour simplifier les formules, nous supposons que nous travaillons sur une image I rectangulaire, de taille $N \times M$. Selon les caractéristiques présentées, nous serons amenés à travailler dans l'espace de couleur classique RGB, dans l'espace HSV (pour teinte H, saturation S et valeur V), ou encore dans l'espace $L^*a^*b^*$ (canal de luminance L^*, et deux canaux de chrominance a^* et b^*). Parfois l'image sera simplement convertie en niveaux de gris, et nous notons ce canal N_G. Par défaut dans notre implémentation, nous travaillons sur des images dont chaque pixel est représenté par 24 bits (3 canaux, 256 valeurs possibles par canal).

Au total, nous proposons 15 caractéristiques distinctes, dont 1 indice de netteté, 2 de texture, 2 d'illumination, 4 de contraste, et 6 de couleur. Différentes photos sont présentées afin d'illustrer chaque catégorie de caractéristiques. Les valeurs de caractéristiques associées aux photos sont les valeurs renormalisées entre 0 et 1 (voir 2.2.2).

3.3.1.1 Mesure de netteté

L'indice de netteté, que nous notons C_1 est calculé en utilisant le procédé décrit dans [Crete et al. 2007]. La valeur de l'indice est d'autant plus proche de 0 que l'image est nette, à l'inverse une image très floue aura une valeur proche de 1. L'idée principale est de comparer l'image originale $\mathcal{I}$, convertie en niveaux de gris (canal N_G), à sa version floutée $\mathcal{B}$ (i.e. filtrée passe-bas). Si les images $\mathcal{I}$ et $\mathcal{B}$ sont très différentes, c'est que $\mathcal{I}$ est nette, sinon $\mathcal{I}$ est floue.

Plus précisément, pour chacune des deux images, deux nouvelles images sont calculées afin de mesurer les variations horizontales et verticales des images $\mathcal{I}$ et $\mathcal{B}$. Nous appelons ces images $\mathcal{DI}_H$, $\mathcal{DI}_V$, $\mathcal{DB}_H$ et $\mathcal{DB}_V$. Elles sont calculées selon les formules suivantes, où (i,j) est le pixel de la ligne j et de la colonne j :

$$\mathcal{DI}_H(i,j) = |\mathcal{I}(i,j) - \mathcal{I}(i,j+1)| \tag{3.1}$$
$$\mathcal{DI}_V(i,j) = |\mathcal{I}(i,j) - \mathcal{I}(i+1,j)| \tag{3.2}$$
$$\mathcal{DB}_H(i,j) = |\mathcal{B}(i,j) - \mathcal{B}(i,j+1)| \tag{3.3}$$
$$\mathcal{DB}_V(i,j) = |\mathcal{B}(i,j) - \mathcal{B}(i+1,j)| \tag{3.4}$$

Il est ensuite calculé la différence entre les variations des images $\mathcal{I}$ et $\mathcal{F}$:

$$\mathcal{V}_H(i,j) = Max(\mathcal{DI}_H(i,j) - \mathcal{DB}_H(i,j),0) \tag{3.5}$$
$$\mathcal{V}_V(i,j) = Max(\mathcal{DI}_V(i,j) - \mathcal{DB}_V(i,j),0) \tag{3.6}$$

Les coefficients suivants permettent de quantifier les variations d'intensité horizontales et verticales sur les deux images $\mathcal{I}$ et $\mathcal{V}$:

$$\mathcal{SI}_H = \sum_{i,j} \mathcal{DI}_H(i,j) \qquad \mathcal{SI}_V = \sum_{i,j} \mathcal{DI}_H(i,j) \tag{3.7}$$
$$\mathcal{SV}_H = \sum_{i,j} \mathcal{V}_H(i,j) \qquad \mathcal{SV}_V = \sum_{i,j} \mathcal{V}_H(i,j) \tag{3.8}$$

Ainsi, $\mathcal{SI}_H$ représente l'intensité totale des variations horizontales de l'image de départ tandis que $\mathcal{SV}_H$ représente le total de la perte de variation d'intensité horizontale entre l'image de départ $\mathcal{I}$ et l'image floutée $\mathcal{B}$. La différence relative entre ces deux valeurs détermine finalement le niveau de flou horizontal f_H de l'image $\mathcal{I}$:

$$f_H = \frac{\mathcal{SI}_H - \mathcal{SV}_H}{\mathcal{SI}_H} \tag{3.9}$$

De même, le flou vertical f_V se calcule par :

$$f_V = \frac{\mathcal{SI}_V - \mathcal{SV}_V}{\mathcal{SI}_V} \tag{3.10}$$

La valeur finale C_1 (pour Caractéristique 1) utilisée sera le maximum entre les valeurs horizontale et verticale : $C_1 = Max(f_H,f_V)$. Cela signifie qu'une image floue uniquement dans une direction (un flou lié à un mouvement par exemple) sera considérée comme floue. Des exemples d'images correspondant à différents niveaux de flou sont donnés sur la figure 3.10. Nous avons choisi d'appliquer cette méthode d'estimation de la netteté plutôt que d'autres méthodes de la littérature, basées sur des décompositions fréquentielles de l'image, car nous avons constaté que cette mesure est plus discriminante.

FIGURE 3.10 – Exemples d'images extraites de la base HFS. De gauche à droite, les valeurs de flou observé sont respectivement 1 (image la plus floue de la base), 0,79, 0,29 et 0 (image la plus nette de la base).

3.3.1.2 Mesures de texture

Afin d'avoir une idée de l'influence des textures dans l'image, nous calculons 2 valeurs. La première correspond simplement à la moyenne des gradients de l'image, notée C_2. L'image des gradients, noté $\mathcal{G}rad$ est obtenue par convolution de l'image en niveaux de gris avec un masque de Sobel. Une forte valeur signifiera que l'image contient de nombreux contours et textures, tandis qu'une faible valeur correspondra à une image uniforme ou floue. Par exemple, un C_1 faible couplé à un C_2 faible suggère une image nette possédant très peu de textures (typiquement un arrière-plan uni).

Dans ce travail, il est également calculé la taille de la surface sur laquelle se concentrent 90% de l'énergie du gradient (caractéristique C_3). Cette information donne un indice de la répartition des informations de l'image. Souvent, une image de bonne qualité esthétique met en avant le visage en proposant un arrière-plan sobre, c'est-à-dire contenant peu de textures. Cela suppose que la surface de l'image dans laquelle les forts gradients sont concentrés est réduite, et la valeur de C_3 est faible. Pour cela, deux tableaux T_H et T_V sont créés. Pour une image de taille $N \times M$, les tableaux sont définis par :

$$\forall i \in 1 \dots N,\ T_H(i) = \sum_{j=1}^{M} \mathcal{G}rad(i,j) \tag{3.11}$$

$$\forall j \in 1 \dots M,\ T_V(j) = \sum_{i=1}^{N} \mathcal{G}rad(i,j) \tag{3.12}$$

Soient $Total_H = \sum_{i=1}^{N} T_H(i)$ et $Total_V = \sum_{j=1}^{M} T_V(j)$ les valeurs obtenues en sommant respectivement les éléments de T_H et de T_V (remarque : $Total_H = Total_V$). Afin de déterminer la taille de la zone contenant 90% des textures, nous calculons pour chaque tableau les indices des éléments pour lesquels la valeur cumulée des précédents éléments dépasse 5% (indice i_5) et 95% (indice i_{95}) du total. La différence $i_{95} - i_5$ correspond à la largeur (pour T_H) ou à la hauteur (pour T_V) du rectangle contenant 90% des textures. La valeur finale retenue sera le rapport $C_3 = \frac{Hauteur \times Largeur}{M \times N}$. Cette valeur sera élevée si les textures sont représentées dans

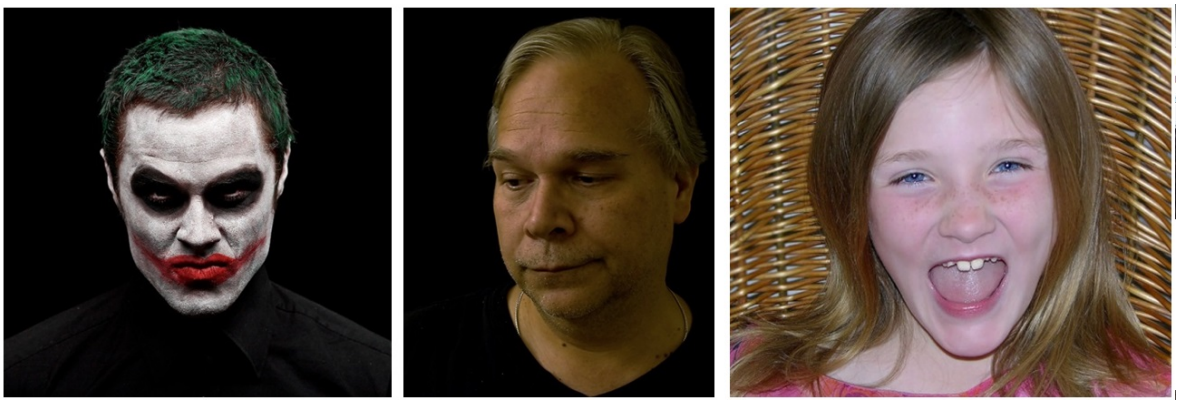


FIGURE 3.11 – Exemples d'images extraites de la base FAVA. Les textures sont uniquement réparties dans le visage (image de gauche et du milieu) et sont réparties dans toute l'image pour la photo de droite. Les valeurs de C_3 sont de 0, 0,3 et de 1 (après normalisation des caractéristiques), de gauche à droite.

toute l'image, et faible si seule une petite partie de l'image contient des textures. Des exemples d'images possédant des répartitions de textures différentes sont proposés sur la figure 3.11.

3.3.1.3 Mesures d'illumination

Pour mesurer l'illumination, nous étudions les deux canaux de luminance que sont les canaux V de l'espace HSV et L^* de l'espace CIE $L^*a^*b^*$. Ces deux canaux sont étroitement corrélés, sans pour autant être égaux. En effet, V est obtenu en prenant la valeur maximale parmi les 3 valeurs de R, G et B de chaque pixel, tandis que L^* s'obtient à l'aide de la formule suivante :

$$L^*(i,j) = \begin{cases} 116 * Y^{1/3} - 16 & \text{pour } Y > 0{,}008856 \\ 903{,}3 * Y & \text{pour } Y \leq 0{,}008856 \end{cases} \text{ où } Y = (0{,}213R + 0{,}715G + 0{,}0721B)/255.$$

Les formules indiquées ici sont celles proposées dans l'implémentation de la bibliothèque OpenCV, que nous utilisons pour l'extraction de caractéristiques. Nous calculons ensuite pour chaque image les moyennes des canaux V et L^*. Ces moyennes sont un indicateur de la luminosité générale de l'image. En les calculant dans des régions particulières (le visage par exemple), il est possible d'obtenir des indications de luminosité locales. Nous avons choisi de conserver les 2 valeurs, car même si les moyennes obtenues sont très semblables, elles ne sont pas identiques et nous obtenons plus d'informations sur l'illumination de l'image. Ces deux moyennes sont notées respectivement C_4 et C_5 pour les canaux V et L^*. Des exemples d'images aux illuminations globales différentes sont données sur la figure 3.12.

3.3.1.4 Mesures de contraste

Quatre indices de contraste sont proposés dans notre modèle. Les deux premiers découlent des mesures d'illumination : C_6 et C_7 représentent respectivement les écarts-types des canaux V et L^*. Ainsi, plus les valeurs de C_6 et de C_7 sont importantes, plus le contraste de l'image

FIGURE 3.12 – Exemples d'images extraites de la base HFS. Les valeurs d'illumination (C_4) sont de 0 (image à gauche, la plus sombre de la base), 0,39, 0,76 et 1 (à droite, image la plus éclairée).

sera élevé. En effet, une valeur importante signifie que les images présentent beaucoup de pixels dont la luminosité est éloignée de la luminosité moyenne.

Deux autres indices de contraste sont pris en compte, et quantifient l'utilisation de la totalité des valeurs de luminance possibles. Pour ces indices, nous considérons le canal L^* uniquement, car nous n'avons pas constaté de différences de performances avec les mêmes valeurs calculées sur le canal V. Le premier de ces indices, noté C_8, est obtenu en appliquant la formule du contraste de Michelson : $C_8 = (L_{max} - L_{min})/(L_{max} + L_{min})$. Si la totalité des valeurs possibles est utilisée ($L_{max} = 255$, $L_{min} = 0$), le contraste de Michelson prend la valeur 1. Dans le cas contraire, où les valeurs de luminance sont toutes très proches, la valeur de C_8 sera très proche de 0.

Le dernier indice calculé, C_9, est la largeur de l'histogramme du canal L^*. Plus précisément, nous calculons la longueur du segment contenant 90% de la masse de cet histogramme. Les valeurs obtenues pour cet indice sont semblables à la formule de Michelson, mais sont cependant beaucoup moins sensibles à la présence de quelques pixels ayant des valeurs de luminance extrêmes.

Pour résumer, nous utilisons dans nos expériences deux indices de la répartition de la luminosité de l'image (C_6 et C_7), et deux indices concernant l'utilisation de l'échelle des valeurs de luminance (C_8 et C_9). Des indices de contraste locaux peuvent ensuite être obtenus en mesurant le contraste dans différentes régions de l'image. Des exemples d'images dont les valeurs de contraste sont différentes sont donnés sur la figure 3.13.

3.3.1.5 Mesures de couleur

Le choix des couleurs présentes dans une photo de visage est un élément important. Des couleurs trop ternes sont généralement synonymes de photos ratées. A l'inverse des photos très colorées mais peu structurées (motifs complexes, peu réguliers) ont tendance à atténuer l'importance du visage dans la photo. La complémentarité des couleurs entre l'avant-plan et

FIGURE 3.13 – Exemples d'images extraites de la base HFS. Les valeurs de contraste sont très faibles à gauche, très élevées à droite (pour des illuminations semblables). Les valeurs de C_6 sont, de gauche à droite : 0,04 (très faible contraste), 0,45, 0,54 et 0,8 (très fort contraste).

l'arrière-plan peut être un atout, mais un arrière-plan uni et peu coloré est souvent présent sur les photographies de professionnels. Des exemples d'images aux couleurs très différentes sont proposés en figure 3.14. Ces exemples montrent que ce n'est pas tant la présence d'une couleur particulière qui influence la qualité esthétique, mais plus leur complémentarité et le nombre de ces couleurs. Les caractéristiques que nous considérons dans nos modèles intègrent donc plutôt des informations de variation (C_{10}, C_{11}), de quantité (C_{12}) et de pureté des couleurs (C_{13}, C_{14} et C_{15}).

Nous définissons les valeurs C_{10} et C_{11} comme étant respectivement les écarts-types des canaux de teinte H et de saturation S. Ces valeurs fournissent des informations sur la distribution des couleurs dans l'image. En effet, un écart-type de H faible indique que les couleurs de l'image sont semblables partout : le visage est peu mis en relief. A l'inverse, des écart-types de H et de S élevés dans des régions particulières (par exemple les yeux, la bouche) sont généralement liés à des images dont les attributs du visage sont saillants : présence de maquillage, yeux grand ouverts, couleurs marquées.

Nous définissons également un critère permettant de quantifier le nombre de couleurs dans l'image. Celui-ci est calculé à partir de l'espace de couleur $L^*a^*b^*$, créé dans le but de faire correspondre les distances mathématiques entre les couleurs aux différences perçues par l'œil humain. Les canaux a^* et b^* permettent de décrire la couleur des pixels, pour lesquels un histogramme h de taille 16×16 est créé. La première dimension correspond à une quantification des valeurs du canal a^*, la seconde du canal b^*. Nous avons choisi un nombre de tranches plutôt faible pour chaque canal de couleur (seulement 16) car certaines des régions dans lesquelles l'histogramme est calculé sont très petites (parfois de l'ordre du millier de pixels). De plus, les pixels dont la valeur L^* est trop faible ou trop élevée ($L^* < 40$ ou $L^* > 240$) ne sont pas pris en compte car leur couleur ne pourra pas être évaluée de manière significative. Notons h_{max} la valeur maximale de l'histogramme et $h(i,j)$ le nombre de pixels tels que les valeurs a^* et b^* appartiennent respectivement aux barres i et j de l'histogramme. Soit $|C|$ le nombre de

FIGURE 3.14 – Exemples de photos pour lesquelles les couleurs sont très différentes. En haut, les couleurs de l'arrière-plan sont vives et contrastent avec celles du visage. Souvent, une seule couleur est présente à l'arrière-plan. En bas à gauche, les couleurs sont liées à la présence d'objets divers qui diminuent la qualité esthétique des clichés. En bas à droite, les couleurs sont ternes et ne contrastent pas avec le visage : la qualité esthétique est donc plus faible également.

couples (i,j) tels que $h(i,j) > \alpha \times h_{max}$. Le nombre de couleurs C_{12} est finalement défini par

$$C_{12} = \frac{|C|}{16 \times 16} \times 100\% \tag{3.13}$$

Dans nos expériences, nous fixons arbitrairement α à 0,001. Des valeurs plus grandes induisent énormément de rejets et donc des valeurs très faibles pour C_{12} dans le cas où nous travaillons avec de petites images, tandis que des valeurs plus petites risquent d'ajouter trop de couleurs représentées par très peu de pixels dans le décompte.

Un autre critère de coloration de l'image est celui proposé par [Hasler et Suesstrunk 2003]. En notant μ_a et μ_b les moyennes respectives des canaux a^* et b^*, ainsi que σ_a et σ_b leurs écart-types, nous calculons la valeur C_{13} à l'aide de la formule :

$$C_{13} = \sqrt{\sigma_a^2 + \sigma_b^2} + 0{,}37\sqrt{\mu_a^2 + \mu_b^2} \tag{3.14}$$

Cette valeur a été utilisée par différents travaux afin de quantifier le caractère coloré d'une image : plus la valeur de C_{13} est élevée, plus l'image (ou la région considérée) est colorée. [Hasler et Suesstrunk 2003] montrent que cet indice est très largement corrélé (coefficient de corrélation autour de 94%) avec le niveau de coloration perçu par des observateurs humains.

Enfin, nous utilisons deux mesures obtenues à partir d'un canal de couleur introduit par [He et al. 2011] dans le but d'atténuer les effets de brouillard sur les photos, et baptisé *Dark channel*. Nous parlerons de canal sombre dans la suite de cet exposé. Il est également utilisé par [Luo et al. 2011] comme critère d'évaluation de la qualité esthétique de photo. Ce canal, noté C_{Sombre}, est construit en calculant le minimum d'intensité local pour tous les canaux, en chacun des points de l'image. Ainsi, pour un pixel (i,j) d'une image I :

$$C_{Sombre}(i,j) = \min_{c \in R,G,B} \left(\min_{(i',j') \in \Omega(i,j)} I_c(i',j') \right) \tag{3.15}$$

où I_c est un canal de I et $\Omega(i,j)$ un voisinage du pixel (i,j). Dans notre étude, ce voisinage est une fenêtre de taille $n \times n$ pixels. Nous posons $n = 11$, qui permet d'obtenir les images présentées sur la figure 3.15. Comme ce canal est un filtre minimal appliqué aux 3 canaux R,G,B, une image floue ou contenant du brouillard aura une valeur moyenne de ce canal sombre plus élevée qu'une image nette. De même, des couleurs ternes engendrent un canal sombre plus élevé. Ce canal englobe ainsi des informations concernant la pureté des couleurs ainsi que la netteté de la photo. Calculer la valeur moyenne ainsi que l'écart type du canal sombre semble donc pertinent ; ces valeurs sont notées C_{14} et C_{15}.

Remarque : Il existe un grand nombre d'images en niveaux de gris dans certaines des bases d'images que nous utilisons. Pour ces images, il n'existe pas d'informations sur la couleur, et seules les informations relatives à la netteté, à la texture, à l'illumination et au contraste sont disponibles. Une solution pour évaluer tout de même les images en noir et blanc est de tripler le canal de niveaux de gris N_G pour obtenir une image à 3 canaux. Les caractéristiques présentées ici, de C_{10} à C_{15}, peuvent ainsi être calculées normalement.

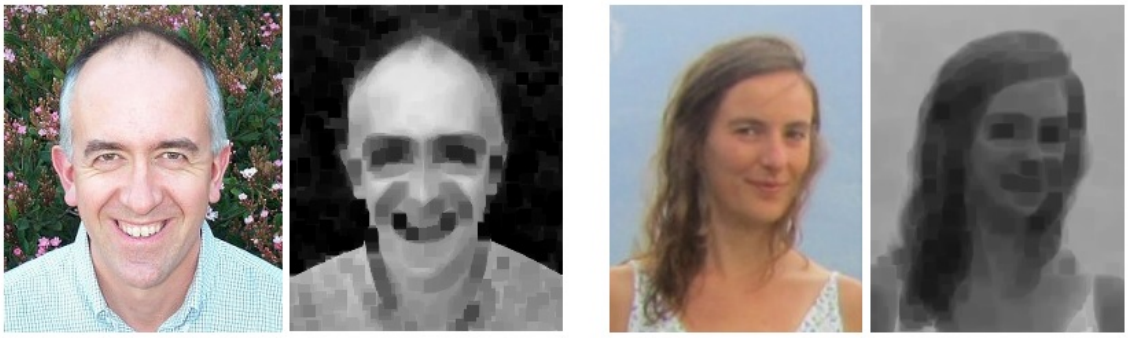


FIGURE 3.15 – Exemples d'images extraites de la base HFS. La valeur du canal sombre est plutôt faible pour l'image de gauche ($C_{14} = 0{,}57$), car l'arrière-plan est constitué d'une couleur vive. L'écart-type du canal sombre est cependant élevé : $C_{15} = 0{,}38$. A l'inverse, la photo de droite présente une valeur moyenne du canal sombre plus élevée ($C_{14} = 0{,}72$), mais un écart-type plus faible ($C_{15} = 0{,}27$) : le visage est peu contrasté par rapport à l'arrière-plan.

3.3.2 Composition de l'image

Jusqu'à présent, nous avons décrit des caractéristiques représentant des informations calculées sur une région quelconque de l'image. Nous proposons maintenant des descripteurs tenant directement compte de la composition de l'image.

Ces descripteurs peuvent par exemple correspondre à des différences de caractéristiques entre deux régions : différence de netteté ou d'éclairage entre le visage et l'arrière-plan, etc. Toutefois, ce type de mesures est déjà implicitement pris en compte par le calcul de caractéristiques sur différentes régions de l'image, et nous n'avons pas observé d'améliorations des performances de nos modèles à l'aide de ces valeurs.

Nous nous limitons finalement à l'utilisation de 3 grandeurs décrivant la position et la taille du visage dans l'image. La première correspond simplement au ratio entre le nombre de pixels de la zone du visage et le nombre de pixels total de l'image. Nous définissons ensuite 2 indices de position, représentant les positions relatives du visage dans l'image, selon les axes horizontaux et verticaux.

Ces 3 indices sont plus ou moins importants selon les contraintes imposées sur les bases d'images. En effet, la plupart des bases que nous considérons (HFS, FAVA, PNF) contiennent des photographies dont la composition est très stricte : la position et la taille des visages ne varient que très peu. Pour cette raison, ces caractéristiques sont uniquement considérées lorsque nous souhaitons comparer les résultats de notre méthode aux résultats de l'état de l'art sur des bases de photographies à la composition plus libre (CUHKPQ, AVA, Flickr).

3.4 Définition des régions du visage considérées pour le calcul des descripteurs

Certains travaux [Datta et al. 2006] proposent de découper l'image en blocs de taille arbitraire (par exemple une grille régulière de taille 3×3), puis de calculer les caractéristiques dans chacun de ces blocs. Cette segmentation présente l'avantage de ne nécessiter aucune information a priori sur le type d'images considéré, et le fait de calculer les caractéristiques dans différentes régions permet d'encoder implicitement des informations spatiales dans les caractéristiques. Il a été montré que le calcul de caractéristiques sur des régions particulières (définies par l'avant-plan ou un visage) améliorent significativement les performances d'estimation de la qualité esthétique [Luo et Tang 2008 ; Kim et Kim 2014]. Dans une photo de visage, le regard de l'observateur est automatiquement attiré par le visage du sujet. Ainsi, lors de la phase d'extraction de caractéristiques, il semble naturel d'extraire des informations particulières dans la région du visage. Toutefois, les modèles actuels n'incluent pas d'informations sur des régions plus fines et cependant très riches en information, telles que la région des yeux et la bouche.

En effet, en regardant une photo de visage, l'observateur humain a tendance à analyser ces régions en priorité car celles-ci contiennent non seulement les informations relatives à la qualité de l'image (une photo dont le visage est flou est généralement une photo ratée), mais contiennent également des informations contextuelles décrites à travers l'expression du sujet (sourire, joie, tristesse). Nous proposons donc de calculer nos attributs de bas niveau sur différentes régions de l'image, et nous décrivons dans cette section les algorithmes permettant la détection de ces éléments (visage, yeux, bouche). Nous montrons ensuite que le calcul d'informations dans ces zones d'intérêt améliore la précision des estimations de la qualité esthétique.

3.4.1 Extraction des différentes zones du visage

Pour la détection du visage et des attributs faciaux (les yeux, la bouche), nous avons choisi d'utiliser l'algorithme proposé par [Viola et Jones 2001] car ce dernier est implémenté dans la plupart des bibliothèques de vision par ordinateur et est très rapide d'exécution. Des algorithmes plus récents et plus robustes existent, mais sont généralement plus complexes à mettre en place car basés sur l'utilisation de réseaux de neurones convolutionnels [Li et al. 2015]. Si l'algorithme que nous utilisons est connu pour ses limitations dans le cas où le visage n'est pas de face ou si une partie du visage est occultée, dans ce document nous ne travaillons qu'avec des visages de face (ou très légèrement orientés) et entièrement visibles.

Dans nos travaux, nous supposons que les photos contiennent forcément au moins un visage. Toutefois, sa position et sa taille ne sont pas connues au départ et influent directement sur la qualité esthétique de l'image : la qualité esthétique d'une photo dépend de règles de composition bien précises. Par exemple, la règle des tiers suggère que le sujet de la photographie soit placé au niveau des intersections des lignes imaginaires divisant l'image en neuf parties

égales. Des visages trop petits ou trop excentrés ont également tendance à atténuer la mise en valeur du sujet.

La bibliothèque OpenCV intègre différentes fonctions permettant de détecter automatiquement et rapidement des objets dans une image grâce à l'algorithme de Viola et Jones [Viola et Jones 2001]. Avant la phase de détection, l'image est convertie en niveaux de gris et le contraste est augmenté à l'aide d'une égalisation d'histogramme. De plus, afin d'accélérer la détection, il est possible de réduire la taille de l'image. Il devient alors difficile de repérer de petits objets, mais la détection d'éléments importants (typiquement un visage dans une photo de visage) est largement accélérée.

3.4.1.1 Algorithme de Viola et Jones

L'algorithme de Viola et Jones permet d'apprendre à reconnaître un objet donné dans une image. Il a été conçu dans un premier temps afin de détecter de manière efficace et en temps réel des visages dans une image, puis a été étendu au cas de n'importe quel objet. L'apprentissage est supervisé et nécessite dans un premier temps deux catégories d'images décrivant respectivement différentes images de l'objet considéré (un visage par exemple) et des images ne contenant pas l'objet considéré. L'apprentissage consiste à distinguer ces deux catégories d'images.

L'algorithme d'apprentissage utilisé est Adaboost, pour "Adaptative Boosting" [Freund et al. 1999]. Celui-ci permet de construire un classifieur fort (performant) à partir d'un ensemble de classifieurs faibles (simples) qui sont dans ce cas les caractéristiques pseudo-Haar développées par Viola et Jones et améliorées par [Lienhart et Maydt 2002]. L'intérêt de ces caractéristiques est leur capacité à être évaluées à coût constant ($\mathcal{O}(1)$) grâce à l'utilisation d'images intégrales (voir figure 3.16). Des fichiers contenant les modèles issus de cet apprentissage pour différents objets (visage, œil droit, œil gauche, nez, bouche) sont inclus dans la librairie OpenCV.

Ensuite, l'image à tester est parcourue à l'aide de fenêtres glissantes de différente taille, ce qui permet de détecter des objets de différente taille. Dans chaque fenêtre, les caractéristiques pseudo-Haar sont calculées puis comparées aux valeurs enregistrées lors de l'apprentissage afin de décider si la fenêtre contient l'objet ou non. La très grande majorité des fenêtres étudiées ne contenant pas l'objet, le processus de décision est accéléré à l'aide d'une cascade de classifieurs. Plus précisément, à chaque étage de la cascade, seul un sous-ensemble des caractéristiques est considéré. Les caractéristiques les plus discriminantes se situent dans les premiers étages de la cascade, de manière à ce que les décisions de rejet soient prises au début de la chaine de traitement. Un schéma résumant le processus de détection d'objet (appliqué à la détection du visage) est donné sur la figure 3.17.

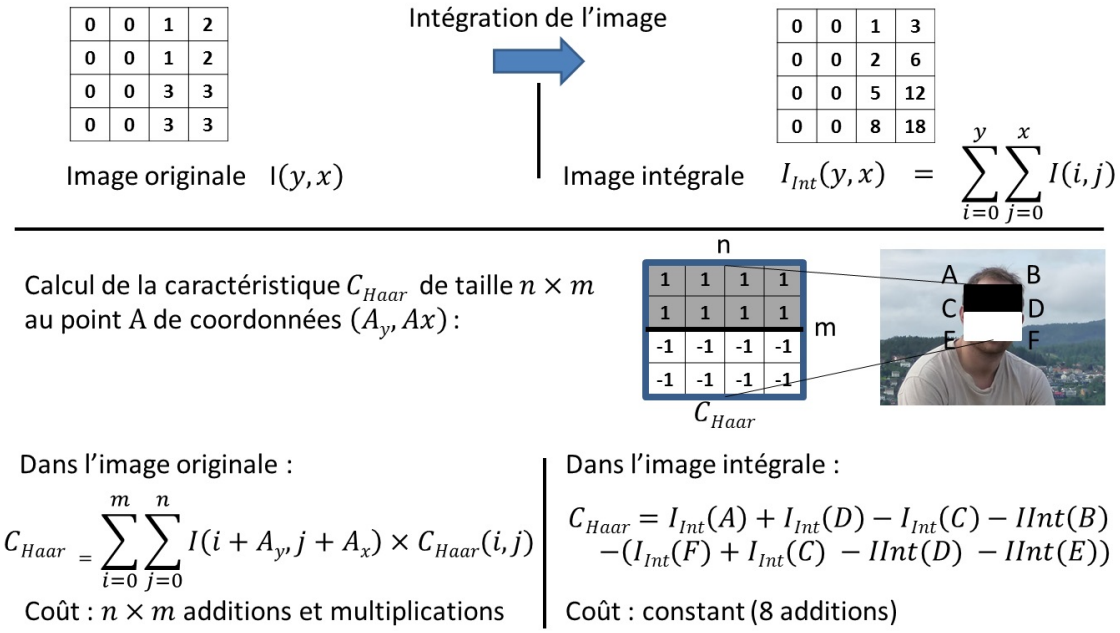


FIGURE 3.16 – Principe de fonctionnement et d'utilisation des images intégrales. Cette transformation permet de calculer à coût constant des sommes de valeurs dans les zones rectangulaires définies par les caractéristiques pseudo-Haar [Lienhart et Maydt 2002].

3.4.1.2 Détail des paramètres

Au total, 5 fichiers différents sont utilisés, contenant les résultats de l'apprentissage des objets suivants : visage, œil gauche, œil droit, nez et bouche. La méthode employée pour détecter chacune de ces parties du visage est la même, si ce n'est que la recherche des attributs faciaux est limitée à la zone définie par l'intérieur de la région du visage.

Différents paramètres ont une influence notable sur le taux de bonnes détections d'un objet donné. Par exemple, la taille minimale de la fenêtre de détection permet d'éviter de faux positifs correspondant à des objets trop petits. La taille minimale de la fenêtre du visage est définie par nos contraintes : la largeur et la hauteur du visage doivent être respectivement supérieures à 1/3 de la largeur et de la hauteur de l'image.

OpenCV permet de définir une taille de fenêtre minimale ainsi qu'une taille maximale : la taille d'un œil ne peut par exemple pas dépasser celle du visage. Les fenêtres contenant les yeux doivent avoir au moins 10% de la hauteur et de la largeur du visage, et les fenêtres décrivant le nez et la bouche doivent être au moins 20% de la hauteur et de la largeur du visage. Il est aussi possible de définir le taux d'agrandissement de la fenêtre lors de chaque parcours de l'image. Un taux trop faible ralentit considérablement l'algorithme, un taux trop élevé ne garantit pas de détecter un visage dont la taille est incluse entre deux tailles de fenêtre. Nous fixons le taux d'agrandissement à 0,1.

Dans le but d'éviter les faux positifs correspondant à des objets ressemblant à un visage,

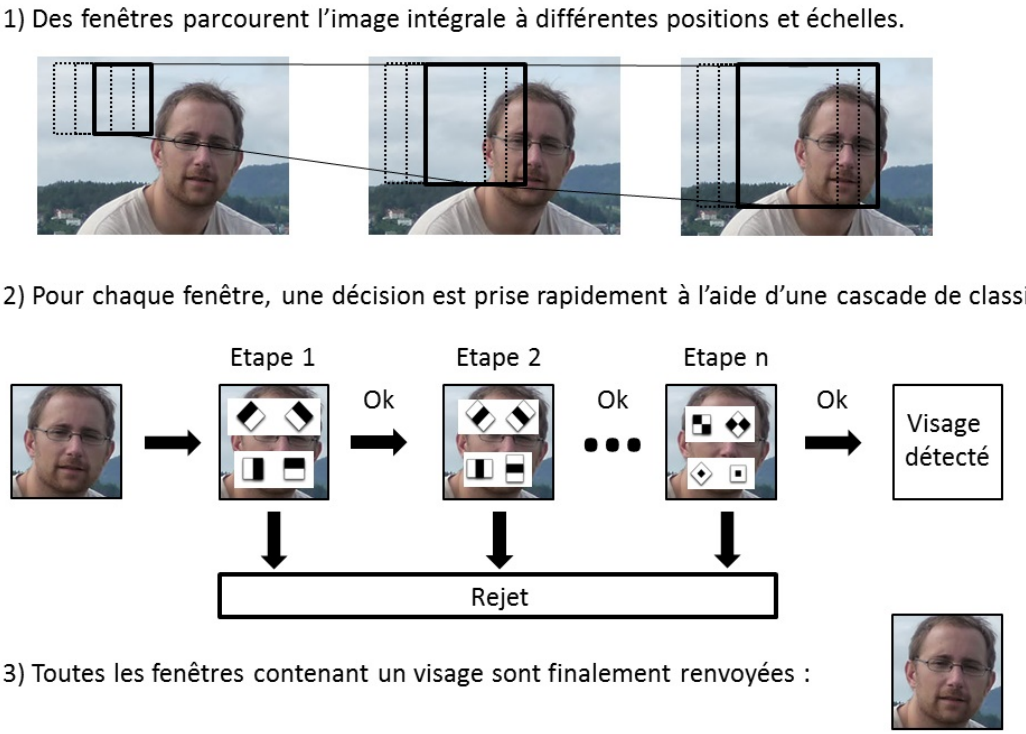


FIGURE 3.17 – Fonctionnement simplifié de l'algorithme de Viola-Jones, appliqué à la détection de visages.

nous pouvons définir le nombre minimal de fenêtres se chevauchant ayant eu une réponse positive. Plus ce nombre est faible, plus le nombre de faux positifs est élevé. À l'inverse, certains visages ne seront pas détectés avec un nombre trop élevé (faux négatifs). Cette valeur est fixée à 2 par défaut dans nos travaux afin de limiter le nombre de faux négatifs (nos images contiennent toujours au moins un visage). Cela signifie qu'en décalant une fenêtre contenant un visage d'un pixel, l'algorithme détecte toujours un visage.

Des exemples de détection sont présentés en figure 3.18 et les performances de détection sont détaillés en 3.4.2.

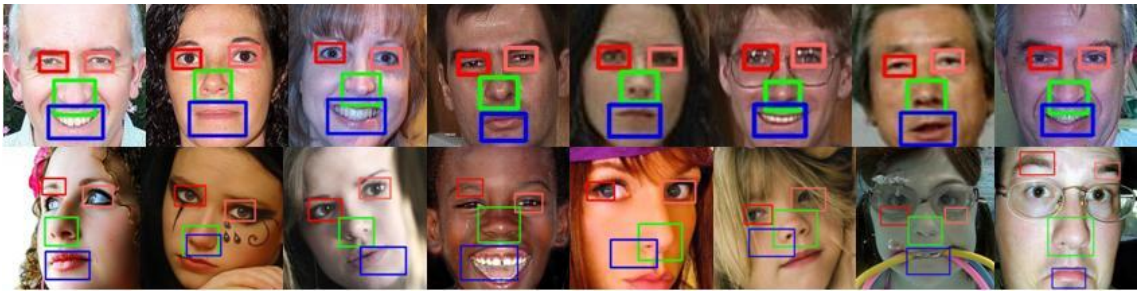


FIGURE 3.18 – Exemples d'objets détectés. La première ligne présente des détections correctes, tandis que dans la seconde ligne, certains attributs ne sont pas toujours correctement détectés.

3.4.1.3 Vérifications supplémentaires

En utilisant les paramètres décris précédemment, il est possible de détecter plusieurs visages dans la photo. Cela peut arriver lorsqu'il y a effectivement plusieurs visages présents dans la photo, ou simplement lorsqu'un objet ressemblant à un visage est détecté. Comme nous n'étudions que les photos contenant un seul visage, nous ne conservons que le plus grand visage détecté. En cas d'égalité, nous choisirons le visage le plus centré.

Il est possible d'effectuer des vérifications après la détection afin d'éviter des situations impossibles ou improbables. Dans notre étude, nous nous limitons aux images dont les visages sont de face. Cela nous permet de détecter les visages et attributs du visage plus précisément. Des détails sur les performances obtenues avec notre détection sont donnés en 3.4.2.

Afin d'être sûr que les visages détectés sont bien de face et qu'il n'y a probablement pas d'erreur de détection, nous vérifions que :

— les yeux sont tous deux sur une même ligne horizontale,
— les yeux sont bien de part et d'autre du nez,
— l'œil gauche est à gauche du visage,
— l'œil droit est à droite du visage,
— les yeux sont au-dessus du nez,
— le nez est dans la partie centrale du visage,
— la bouche se situe sous le nez,
— la bouche est dans la partie basse du visage.

Il arrive que le visage ou l'un de ses différents attributs (les yeux, le nez, la bouche) ne soient pas correctement détectés (faux négatif). Dans ce travail, seules les photos dont les détections sont possibles sont traitées. En effet, nous ne tenons pas compte dans nos expériences des photos dont la détection de tous les éléments ne s'effectue pas correctement.

Les détections erronées (faux positifs) sont rares, mais peuvent arriver malgré les vérifications, ce qui a pour désavantage de fausser les résultats lors des analyses et calculs de caractéristiques.

3.4.2 Performances de la détection des attributs

Afin d'évaluer correctement les photos de visage à l'aide de la méthode proposée dans ce chapitre, il est nécessaire d'obtenir de bons taux de détection des différents attributs du visage. Nous avons cherché à quantifier les performances de la détection des attributs, et avons effectué des tests sur la base de photos de visage LFW [Huang et al. 2007], très utilisée dans le cadre de la reconnaissance faciale et contenant plus de 13000 photos de visage. Pour cette base, la détection du visage ne devrait pas être un problème car une des contraintes de la base correspond justement au fait que les visages doivent pouvoir être détectés à l'aide l'algorithme de Viola et Jones. Nous vérifions ainsi sur cette base que notre algorithme fonctionne correctement et que les différents attributs (les yeux, le nez, la bouche) sont bien détectés.

TABLEAU 3.2 – Nombre d'attributs manquants sur le total de 13233 photos de visage de la base LFW, en utilisant l'algorithme de Viola-Jones.

Attribut	Visage	Œil droit	Œil gauche	Nez	Bouche
Détections manquantes	24	754	611	601	176

Faux négatifs

Nous observons un taux de visages correctement détectés supérieur à 99,8% (24 visages non détectés sur 13223). Les quelques visages qui ne sont pas correctement détectés (mauvaise position du visage, ou pas de visage détecté) proviennent du fait que les paramètres de l'algorithme que nous utilisons ne sont pas forcément identiques à ceux utilisés pour la création de la base LFW.

Nous utilisons la base LFW essentiellement dans le but d'évaluer le taux de détection correcte des différents attributs faciaux que sont le nez, la bouche ou les yeux. Le nombre d'attributs non détectés est recensé dans le tableau 3.2. Nous obtenons un taux de faux négatifs de l'ordre de 2% pour la détection de la bouche, et d'environ 5% pour les autres attributs. Au total, nous observons 734 images (soit 6%) pour lesquelles il y a au moins deux attributs non détectés. Lorsqu'un ou plusieurs attributs manquent, il est possible d'estimer leur position à partir de celle des autres : le nez se situe entre les yeux et la bouche, la bouche se situe sous le nez dans la zone du visage, la position d'un œil manquant s'obtient à l'aide de la taille du visage et de l'autre œil, etc.

Le taux de faux négatifs augmente significativement lorsque d'autres bases d'images sont considérées. Typiquement, sur une base de 500 images contenant des visages récupérées sur le site [*Flickr*], nous ne détectons que 455 visages (91%), parmi lesquels se trouvent 4 faux positifs. Cela s'explique par la petite taille des visages (en nombre de pixels) sur certaines images ou encore par la difficulté pour l'algorithme à trouver des visages légèrement orientés ou occultés.

Faux positifs

Différents types de détections erronées sont dues à la mauvaise estimation de la position d'un attribut. Une des erreurs les plus fréquemment effectuées par l'algorithme est la confusion entre un œil et son sourcil (même type de forme, mêmes emplacements possibles, cf. figure 3.18). Lorsque la boite englobant le visage détecté est approximativement localisée, le bas du visage (au niveau du menton) est parfois confondu avec la région de la bouche. Dans ce cas, il arrive que les narines soient confondues avec la bouche, ou que le nez soit situé trop haut dans le visage. La bouche peut éventuellement être confondue avec une moustache. Ces exemples montrent la difficulté d'utilisation de ce type d'algorithme. Toutefois, ces erreurs sont plutôt rares (entre 1 et 10% des cas selon les bases de données) et n'introduisent pas de grandes erreurs dans nos prédictions de qualité esthétique : le fait de calculer des caractéristiques à plusieurs niveaux permet d'introduire de la redondance dans les valeurs, ce qui induit une certaine robustesse aux erreurs de placement dues à l'algorithme de Viola et Jones. Le fait

de procéder à toutes les vérifications que nous avons décrites rend également le nombre de visage détectés à tort très faible. Ainsi, si nous détectons un visage et que nous ne pouvons détecter ses différents attributs, il est probable que le visage détecté soit un faux positif, et nous rejetons le visage détecté.

Temps d'exécution

Le temps d'exécution est un indice de performance important lorsque l'application doit tourner en temps réel, ou tout simplement s'il y a un très grand nombre d'images à traiter. L'algorithme de Viola-Jones a été conçu afin de justement répondre à cet objectif de rapidité, et c'est entre autre pour cela que nous utilisons cet algorithme. Sur un ordinateur de bureau classique, il est possible de traiter entre 5 et 30 images par seconde (selon la taille de l'image et les paramètres considérés). Si le visage est suffisamment grand dans l'image, la taille de l'image importe peu car il est possible de procéder à un sous-échantillonnage de l'image avant l'application de l'algorithme sans diminuer significativement les performances. Nous montrons dans la dernière section de ce chapitre qu'il est possible d'analyser précisément la qualité esthétique d'images en temps réel, sans pour autant diminuer significativement les performances d'évaluation.

3.4.3 De la détection à la définition des régions considérées

Après avoir calculé les positions des différents attributs faciaux dans l'image, nous définissons au total 4 régions dans l'image. Ces régions sont utilisées tout au long de ce chapitre, car nous extrayons les caractéristiques décrites en 3.3 sur chacune d'entre elles. Elles sont directement issues des différents attributs détectés par l'algorithme de Viola-Jones, et définies sur la figure 3.19.

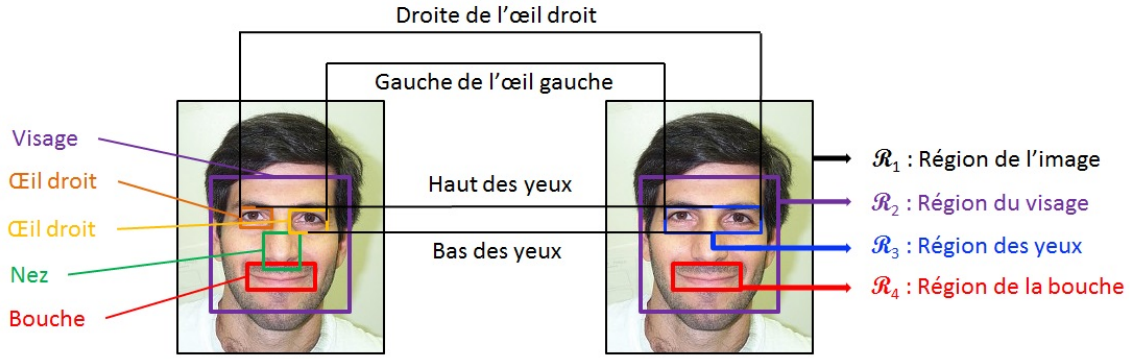


FIGURE 3.19 – Régions détectées par l'algorithme de Viola et Jones (à gauche), et régions utilisées dans le cadre de ce travail (à droite).

Les 4 régions définies correspondent à différents niveaux de l'image. La première région, $\mathcal{R}_1$, correspond à la région définie par l'image entière. Le calcul de caractéristiques sur l'image

entière donne un premier aperçu de la qualité esthétique globale de l'image. Une seconde région, $\mathcal{R}_2$, est définie par la région du visage telle que détectée par l'algorithme de Viola-Jones. Il est ainsi possible de calculer des caractéristiques sur cette région particulière, de manière à extraire des informations adaptées à l'évaluation de photos de visage. Les informations obtenues dans les régions $\mathcal{R}_1$ et $\mathcal{R}_2$ sont déjà prises en compte dans la plupart des travaux sur la qualité esthétique de photos contenant des visages. Dans ce travail, nous analysons l'image à un degré de précision plus élevé en intégrant également les régions $\mathcal{R}_3$ et $\mathcal{R}_4$, correspondant respectivement à la région des yeux et de la bouche (voir figure 3.19). L'extraction d'informations pertinentes dans ces deux régions ainsi que les conclusions des résultats qui en découlent sont les principales contributions de ce chapitre. La région définie par le nez n'est pas utilisée dans nos modèles car celle-ci contient moins d'informations (textures, couleurs...) que les autres régions du visage.

3.5 Étude de la pertinence des descripteurs et des régions

Certaines des caractéristiques considérées peuvent n'être pertinentes que dans une région particulière. Par exemple, les photos de visage de bonne qualité esthétique présentent généralement un arrière-plan flou et des traits nets sur le visage. Calculer le niveau de flou global de l'image n'est donc pas suffisant. A l'inverse, calculer chaque caractéristique dans chaque région de l'image peut diminuer les performances globales de l'algorithme par l'introduction de données redondantes ou peu pertinentes.

Nous extrayons au total 15 caractéristiques sur 4 régions différentes (l'image entière R_1, le visage R_2, les yeux R_3, la bouche R_4), soit 60 valeurs réelles décrivant chaque image. Afin de désigner les 60 valeurs, nous utilisons pour la suite de ce chapitre la notation (C,R), où C désigne l'une des 15 caractéristiques (de C_1 à C_{15}) et R l'une des 4 régions (de R_1 à R_4). Trouver les couples (C,R) les plus discriminants pour l'estimation de la qualité esthétique présente plusieurs avantages. Il est alors possible de calculer moins de caractéristiques peu pertinentes, ce qui réduit le coût en termes de temps de calcul, tout en augmentant la précision du modèle.

Dans ce travail, les 60 couples (C,R) sont classés selon l'algorithme Relief, décrit dans le chapitre 2. Cet algorithme renvoie une valeur réelle sur la capacité de chaque couple à distinguer des images similaires dont les scores de qualité esthétique sont différents.

Les caractéristiques étant analysées par l'algorithme Relief avant la phase d'apprentissage, cette étape est entièrement indépendante du choix de l'algorithme d'apprentissage. Les expériences effectuées dans cette section sont donc uniquement faites à l'aide de l'algorithme SVM sur les bases HFS et FAVA, qui respectent nos contraintes sur la taille et la position du visage dans la photo.

TABLEAU 3.3 – Valeurs de $\mathcal{R}(C)$ calculées pour chaque caractéristique, pour les problèmes de classification (-C) et de régression (-R). Les valeurs indiquées dans le tableau sont multipliées par 100 par rapport aux valeurs réelles. Les valeurs les plus élevées sont indiquées en gras.

Base	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}
HFS-C	**42**	**33**	**16**	12	13	12	13	10	11	12	5	14	14	**16**	14
HFS-R	**23**	**19**	8	10	12	13	2	14	12	7	1	**20**	10	**19**	**15**
FAVA-C	**7**	**7**	**6**	**3**	1	0	**3**	0	0	1	0	2	0	2	**3**
FAVA-R	**19**	8	**20**	14	14	**15**	11	**15**	12	8	**16**	14	14	12	12

3.5.1 Influence des caractéristiques

L'algorithme Relief est utilisé pour classer les 60 couples (caractéristique, région). Afin d'analyser précisément la contribution de la caractéristique C_i uniquement (sans l'influence des régions), nous sommons les contributions des différents couples (C_i, région) :

$$\mathcal{R}(C_i) = \sum_{j=1}^{4} \mathcal{R}(C_i,R_j) \tag{3.16}$$

où $\mathcal{R}(C_i,R_j)$ est la valeur obtenue par l'algorithme Relief pour le couple (C_i,R_j). Pour chacune des caractéristiques C_i, nous reportons ainsi la valeur $\mathcal{R}(C_i)$ dans le tableau 3.3. Nous montrons dans ce tableau les valeurs obtenues pour chaque caractéristique, dans les cas de classification à 2 classes et de régression, pour les bases de données HFS et FAVA. Nous prenons ces deux bases en exemple car les types de photos qui y figurent sont très différentes, et cette différence est visible dans le tableau 3.3 : les caractéristiques discriminantes pour une base d'images ne le sont pas forcément pour une autre. Il apparaît tout de même clairement que les indices liés à la netteté de l'image sont les plus discriminants ($\mathcal{R}(C_1)$, $\mathcal{R}(C_2)$ et $\mathcal{R}(C_3)$ sont souvent parmi les 5 valeurs les plus élevées), ce qui n'est pas surprenant car une image floue (notamment lorsque le visage est flou) est quasiment toujours associée à une image de mauvaise qualité esthétique.

Il ressort également de ces tableaux que dans certains cas, des caractéristiques ne sont pas du tout discriminantes (valeurs très proches de 0), mais qu'aucune de nos caractéristiques n'est pas du tout discriminante pour tous les cas. Cela confirme la pertinence de chacune des caractéristiques que nous conservons dans notre modèle global. Notons que les valeurs renvoyées par l'algorithme Relief ne sont pas normalisées dans le tableau 3.3. Les différences significatives entre les ordres de grandeur des valeurs obtenues pour HFS et FAVA (surtout pour la classification) sont un indicateur de la difficulté du problème. Typiquement, classer les photographies de FAVA en 2 catégories est un problème dont la difficulté essentielle provient de la présence d'un très grand nombre de photos moyennes. Il n'existe alors pas de caractéristiques aussi discriminantes que pour la base HFS.

Afin de vérifier que les valeurs renvoyées par Relief indiquent effectivement la capacité à discriminer des images dont les classes sont différentes (ou dont les scores sont différents), nous analysons les performances de classification (ou de régression) obtenus en utilisant les carac-

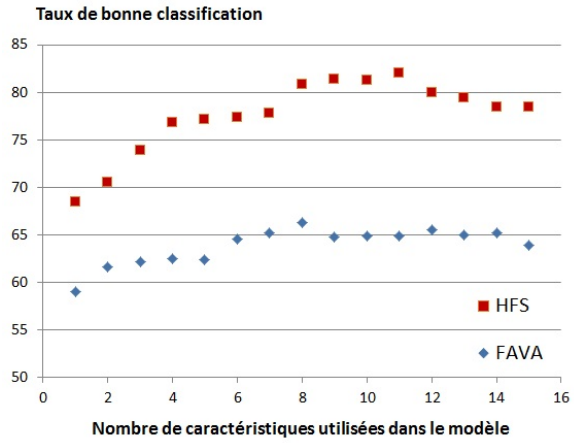


FIGURE 3.20 – Performances des modèles de classification lorsque seules les meilleures caractéristiques sont considérées.

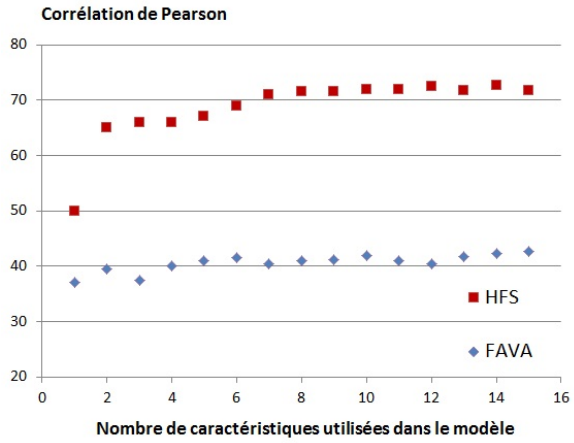


FIGURE 3.21 – Performances des modèles de régression lorsque seules les meilleures caractéristiques sont considérées.

téristiques dont les valeurs renvoyées par l'algorithme Relief sont les plus élevées. Rappelons que dans cette section, lorsque nous étudions une caractéristique, celle-ci est calculée dans les 4 régions de l'image : chaque caractéristique est ainsi représentée par 4 valeurs.

La figure 3.20 représente les performances de classification (en ordonnée) pour les bases HFS et FAVA obtenues en gardant les n (axe des abscisses) caractéristiques les plus discriminantes (indiquées par les valeurs du tableau 3.3). Nous voyons que pour les deux bases d'images, en utilisant un nombre réduit de caractéristiques, il est possible d'obtenir des performances supérieures à celles obtenues avec l'ensemble des caractéristiques ($n = 15$). Cela signifie que l'introduction d'un trop grand nombre de caractéristiques n'améliore pas forcément les performances du modèle. Une explication possible est le fait que pour effectuer de la classification à 2 classes, seul un petit nombre de caractéristiques est nécessaire, et l'introduction d'un grand nombre d'informations risque de générer un sur-apprentissage lié au grand nombre de photos "moyennes", difficiles à classer. A l'inverse, un petit nombre de caractéristiques pertinentes permet de ne tenir compte que des informations importantes, fournies par les images extrêmes.

Nous n'observons pas le même phénomène pour la régression (figure 3.21), pour laquelle réduire le nombre de caractéristiques n'améliore pas les performances. En effet, attribuer un score à une image est un problème plus complexe, pour lequel utiliser plus d'informations est utile. Que ce soit pour les problèmes de classification ou de régression, il est possible de réduire fortement la complexité du modèle (retirer 5 ou 10 caractéristiques) sans pour autant réduire les performances. Toutefois, comme l'indique le tableau 3.3, les caractéristiques à retirer sont dépendantes du problème considéré (classification ou régression).

TABLEAU 3.4 – Performances de classification (Cl.) et de régression (Rg.) obtenues pour chaque région, mesurées respectivement par le taux de bonne classification (T_{BC}) et la corrélation de Pearson (R), en pourcents.

	R_1		R_2		R_3		R_4		R_1,R_2		$R_1 \ldots R_4$	
Base	Cl.	Rg.	Cl.	Rg.	Cl.	Rg.	Cl.	Rg.	Cl.	Rg.	Cl.	Rg.
HFS	74	54	74	61	**77**	**62**	75	60	74	62	**79**	**72**
FAVA	64	34	65	40	**66**	**44**	62	30	65	38	**65**	**42**

3.5.2 Influence des régions

Pour mesurer la pertinence des régions que nous avons définies, nous effectuons de la prédiction à partir de l'ensemble des caractéristiques (de C_1 à C_{15}) sur chacune des 4 régions prises séparément. La plupart des modèles de l'état de l'art ne considèrent que 2 régions (l'image entière et le visage, respectivement R_1 et R_2), et nous souhaitons démontrer l'intérêt de calculer des caractéristiques dans les régions R_3 et R_4. Le tableau 3.4 présente les performances de classification et de régression obtenues pour les deux bases HFS et FAVA, pour chaque région.

Quels que soient le problème et la base considérés, nous observons que la région définie par les yeux (R_3) est celle dans laquelle le calcul des caractéristiques est le plus pertinent. Ceci s'explique par le fait que cette région contient l'essentiel des informations nécessaires à la réussite d'une photo de visage : avant-plan (qui inclut les yeux) net et contrasté, yeux ouverts et visibles (couleurs et contrastes corrects dans la région), etc. Ces informations ne suffisent pas à établir un modèle complet et précis, mais permettent d'avoir une première idée de la qualité esthétique globale de l'image. Ce résultat est doublement intéressant car non seulement le calcul de caractéristiques sur la zone des yeux permet d'améliorer les performances globales de classification, mais il devient également possible d'effectuer des calculs très rapides car cette région est très petite rapportée à la taille de l'image.

Il existe des cas où l'extraction de caractéristiques uniquement dans la région des yeux est plus performante que l'utilisation conjointe de toutes les régions (par exemple pour FAVA, voir le tableau 3.4). Ceci peut s'expliquer par le fait que l'ajout d'un trop grand nombre de caractéristiques introduit du bruit et des informations redondantes. En utilisant un algorithme de sélection des caractéristiques, il est possible de limiter ce phénomène en tenant compte uniquement des informations pertinentes (par exemple sur la netteté de l'image) sur des régions particulières (par exemple les yeux).

3.5.3 Bilan : caractéristiques et régions pertinentes

Nous avons montré que certaines caractéristiques sont plus pertinentes que d'autres, et qu'il en est de même pour les régions. En utilisant uniquement les combinaisons de caractéristiques et de régions pertinentes, il est possible d'améliorer les performances des modèles. Nous reproduisons ainsi l'expérience proposée en 3.5.1 en conservant cette fois les couples (C,R) pour lesquelles les valeurs renvoyées par l'algorithme Relief sont les plus élevées. Toujours en

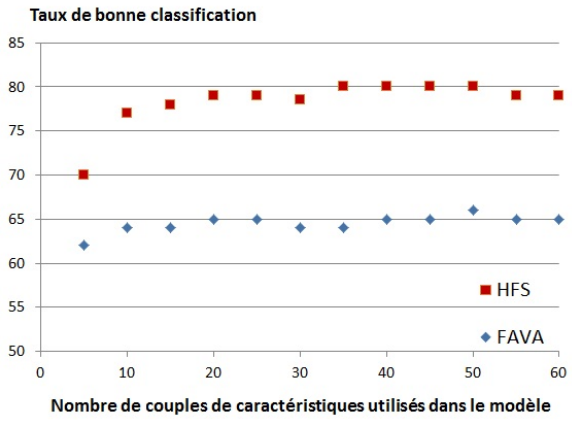


FIGURE 3.22 – Performances des modèles de classification lorsque seuls les meilleurs couples (C,R) sont considérés (voir tableau 3.5).

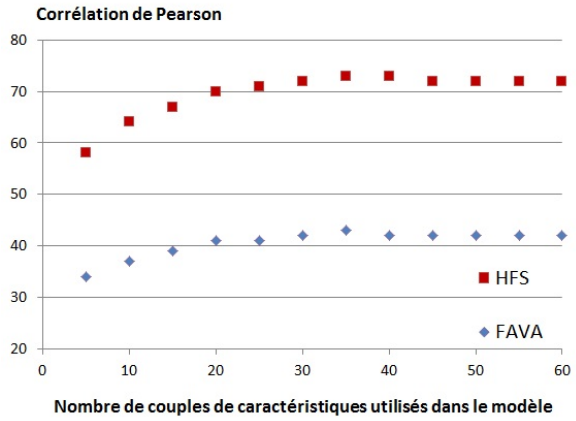


FIGURE 3.23 – Performances des modèles de régression lorsque seuls les meilleurs couples (C,R) sont considérés (voir tableau 3.5).

TABLEAU 3.5 – Caractéristiques et régions correspondant aux 5 couples les plus discriminants (valeurs de $\mathcal{R}(C,R)$ les plus élevés) pour les bases HFS et FAVA, pour les problèmes de classification (-C) et de régression (-R). Les caractéristiques sont ordonnées de gauche à droite : le couple 1 est celui dont la valeur de Relief est la plus élevée.

Base	Couple 1	Couple 2	Couple 3	Couple 4	Couple 5
HFS-C	(C_1,R_3)	(C_1,R_4)	(C_1,R_2)	(C_1,R_1)	(C_3,R_2)
HFS-R	(C_3,R_3)	(C_6,R_4)	(C_1,R_3)	(C_8,R_3)	(C_1,R_1)
FAVA-C	(C_1,R_2)	(C_3,R_3)	(C_1,R_3)	(C_2,R_3)	(C_3,R_1)
FAVA-R	(C_1,R_2)	(C_7,R_4)	(C_3,R_3)	(C_1,R_3)	(C_3,R_1)

utilisant l'algorithme SVM, nous obtenons les figures 3.22 et 3.23, décrivant respectivement les performances de classification et de régression pour les deux bases. Sur ces figures, il apparaît que dans chaque configuration, il est possible d'obtenir les mêmes performances en ne conservant que la moitié des caractéristiques. En utilisant légèrement plus de caractéristiques (autour de deux tiers du jeu complet), il est même possible de légèrement dépasser les performances du modèle global.

Enfin, notons qu'un des intérêts de l'algorithme Relief est l'analyse des couples (C,R) discriminants. Rappelons que le tableau 3.3 présente les caractéristiques les plus discriminantes pour chaque problème et chaque base, et le tableau 3.4 présente les régions dans lesquelles il est le plus intéressant de calculer les caractéristiques. En étudiant plus précisément les couples (C,R) les plus discriminants, nous obtenons directement à l'aide de l'algorithme Relief le tableau 3.5. Ce tableau résume, pour chaque base et chaque problème, les 5 couples (C,R) pour lesquelles la valeur renvoyée par l'algorithme Relief est la plus élevée. Le tableau 3.6 explicite les couples décrits dans le tableau 3.5.

Nous remarquons encore une fois que les caractéristiques les plus discriminantes sont pour

TABLEAU 3.6 – Résumé du tableau 3.5. Pour chaque base d'images, un bref descriptif des descripteurs les plus pertinents est donné.

Base	Caractéristiques et régions les plus pertinentes
HFS-C	Netteté des yeux, de la bouche, du visage, de l'image entière.
HFS-R	Netteté, texture et contraste des yeux. Netteté de l'image entière.
FAVA-C	Netteté du visage. Netteté et texture des yeux, Textures de l'image entière.
FAVA-R	Netteté du visage. Netteté et contraste des yeux. Textures de l'image entière.

la très grande majorité des informations liées à la netteté (C_1,C_2,C_3) calculées sur des régions associées au visage (R_2,R_3,R_4). En observant ce tableau, nous pouvons par exemple remarquer la présence de la caractéristique décrivant la répartition des gradients dans l'image (C_3, voir figure 3.11) dans les caractéristiques discriminantes de la base FAVA. Typiquement, des photos de bonne qualité esthétique présentent un visage net dans une partie de la photo, tandis que l'arrière-plan est généralement très peu texturé. Ce critère est largement pris en compte par les photographes professionnels dont les clichés sont présents dans FAVA, tandis que les photos de la base HFS ne tiennent pas du tout compte de cet aspect (photos amateurs). Ceci suggère que selon l'objectif visé (trier des photographies de professionnels ou trier des photos privées d'amateurs), il est possible de prendre automatiquement en compte ou non certaines caractéristiques. Enfin, notons que les informations de contraste dans les régions définies par les yeux ou la bouche ((C_8,R_3) et (C_6,R_4)) sont également discriminantes. Les informations de contraste et de netteté, décrivant par ailleurs également la qualité générale de l'image, semblent donc être les éléments les plus importants permettant d'évaluer la qualité esthétique d'une photo de visage.

Il ressort également de ces expériences que pour deux bases d'images différentes ou deux problèmes (classification ou régression) différents, les caractéristiques pertinentes ne sont pas exactement les mêmes. Nous n'avons en effet pas réussi à obtenir de jeu de caractéristiques pertinentes efficaces dans toutes les situations. Cela signifie que lorsque nous considérons une nouvelle bases d'images, nous calculons la totalité des 60 valeurs, puis nous appliquons l'algorithme Relief afin de déterminer quels sont les couples de valeurs pertinentes pour les données considérées.

3.6 Étude de la pertinence des algorithmes d'apprentissage supervisé

Nous venons de montrer la capacité de l'algorithme Relief à sélectionner les informations pertinentes dans nos données, ce qui permet d'améliorer légèrement les performances de classification et de régression, en supprimant les caractéristiques peu pertinentes des modèles. Nous avons jusqu'à présent considéré un unique algorithme d'apprentissage : l'algorithme SVM. Dans cette section, nous cherchons à analyser plus précisément l'intérêt de chaque algorithme d'apprentissage, ainsi que les performances de la méthode de fusion des scores proposée dans

TABLEAU 3.7 – Performances des différents algorithmes sur les bases HFS et FAVA, pour les problèmes de classification (-C) et de régression (-R). Les résultats correspondent aux taux de bonne classification (T_{BC} en %) pour la classification, et à la corrélation de Pearson (R en %) pour la régression. Les écarts-types des performances sont également fournis dans le tableau.

Base	SVM	ANN	RF	GBT	LSF
HFS-C	$78{,}4 \pm 1{,}1$	$75{,}8 \pm 1{,}7$	$76{,}8 \pm 1{,}0$	$76{,}1 \pm 1{,}4$	**79,9±0,9**
HFS-R	$71{,}2 \pm 1{,}2$	$59 \pm 4{,}5$	$69{,}0 \pm 1{,}2$	$67{,}3 \pm 0{,}8$	**72,9±1,0**
FAVA-C	$64{,}3 \pm 0{,}9$	$61{,}9 \pm 1{,}3$	$65{,}5 \pm 1{,}4$	$66{,}2 \pm 0{,}9$	**66,8±1,3**
FAVA-R	$41{,}9 \pm 1{,}5$	$41{,}5 \pm 2{,}5$	$46{,}7 \pm 1{,}1$	$48{,}3 \pm 1{,}1$	**49,5±0,8**

le chapitre précédent.

Nous étudions ici les performances de chaque algorithme, pour différentes bases de données et différents types de problèmes. Nous présentons ainsi les avantages et les faiblesses de chacun des algorithmes, et introduisons les raisons qui nous incitent à fusionner les résultats de différents algorithmes afin d'améliorer les performances globales de nos modèles. Dans les expériences proposées dans cette section, nous appliquons dans un premier temps l'algorithme Relief, nous conservons les caractéristiques les plus pertinentes (nous avons observé en 3.5.3 que 40 couples (C,R) sont suffisants), puis nous comparons les résultats de chaque algorithme.

Le tableau 3.7 montre que les algorithmes produisant les meilleures performances ne sont pas les mêmes pour toutes les bases d'images. Typiquement, SVM permet d'obtenir les meilleures performances sur la base HFS, mais l'algorithme GBT est plus efficace sur FAVA. Toutefois, les résultats varient peu d'un algorithme à l'autre : les intervalles de confiance associés aux performances de chaque algorithme se recoupent souvent. Ainsi, plutôt que d'associer chaque base d'images et chaque problème à l'algorithme le plus performant, nous proposons de fusionner les résultats de chaque algorithme selon la méthode proposée au chapitre 2.

Les résultats obtenus par fusion des différents algorithmes d'apprentissage (LFS) sont également donnés dans le tableau 3.7. Non seulement cette méthode permet d'améliorer les performances de classification et de régression, mais aussi de réduire l'écart-type des performances ce qui assure une meilleure stabilité des résultats. Dans certains cas (par exemple pour la classification des images de la base FAVA), les performances ne sont pas particulièrement améliorées par rapport aux performances d'un seul algorithme (GBT dans ce cas). Cela s'explique par le fait que lorsqu'un des algorithmes est significativement plus performant que les autres, il n'est plus forcément intéressant de prendre en compte d'autres algorithmes, pour lesquels les erreurs de classification sont plus fréquentes. C'est pour cette raison que nous avons introduit le coefficient p dans la formule de fusion des résultats (voir équation 2.23), dont le rôle est d'amplifier les différences de poids entre deux algorithmes dont les performances sont inégales.

3.7 Estimation de la qualité esthétique et validation du modèle

Dans cette section nous cherchons à évaluer la qualité esthétique des photographies contenues dans les bases HFS, FAVA et PNF, qui respectent nos contraintes sur la taille et la position du visage dans le photo. Nous extrayons les caractéristiques définies sur les différentes régions de la photo (l'image entière, le visage, les yeux, la bouche) puis nous appliquons les algorithmes définis au chapitre 2.

Pour chaque expérience, nous choisissons de ne conserver que 40 des 60 couples (Caractéristique, Région) pour lesquels la valeur renvoyée par l'algorithme Relief est la plus élevée. Nous avons en effet observé sur les figures 3.22 et 3.23 que les performances de prédiction sont ainsi optimales. Pour l'apprentissage supervisé, nous ne considérons que l'algorithme LSF, pour lequel nous avons obtenu les meilleurs résultats dans le tableau 3.7.

Pour chaque résultat proposé, nous effectuons une validation croisée à 10 groupes que nous répétons 10 fois afin de s'assurer de la stabilité des performances (voir le protocole défini au chapitre 2).

3.7.1 Classification binaire

Sélection d'images aux scores extrêmes - CUHKPQ

Nous cherchons dans un premier temps à montrer que nos caractéristiques et nos algorithmes d'apprentissage sont suffisamment discriminants pour distinguer deux catégories d'images aux scores de qualité esthétique extrêmes. En effet, dans le cadre d'une application réelle, il est intéressant de pouvoir distinguer les photographies très esthétiques, susceptibles d'intéresser l'utilisateur, des photographies ratées. Pour cela, nous effectuons des expériences sur la base CUHKPQ. Celle-ci présente des images évaluées comme étant soit de très bonne, soit de très mauvaise qualité esthétique. Nous étudions les performances de classification en calculant l'aire sous la courbe ROC. Les résultats peuvent être observés sur les figures 3.24 et 3.25. Rappelons que les régions R_1, R_2, R_3, R_4 correspondent respectivement à l'image entière, à la région du visage, des yeux, de la bouche.

Nous observons tout d'abord sur la figure 3.24 que lorsque nous ajoutons des informations locales (calcul de caractéristiques sur le visage), les performances de classification augmentent significativement. Ajouter des informations issues des régions correspondant aux yeux et à la bouche permet encore d'améliorer nettement les performances par rapport aux méthodes ne s'intéressant qu'à une segmentation visage / arrière-plan.

En observant la figure 3.25, nous pouvons étudier la capacité de chaque catégorie de caractéristiques (netteté et textures, illumination, contraste et couleurs) à discriminer les deux catégories de photos. En dehors des valeurs d'illumination, chaque catégorie prise séparément est déjà particulièrement discriminante (aires sous les courbes supérieures à 0,9). Les plus faibles performances des caractéristiques d'illumination s'expliquent par le fait que nous n'uti-

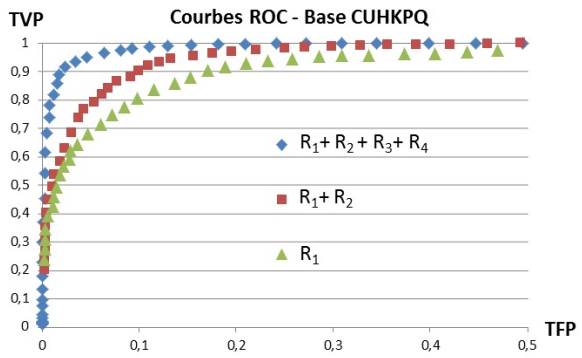


FIGURE 3.24 – Influence des zones de calcul des caractéristiques.

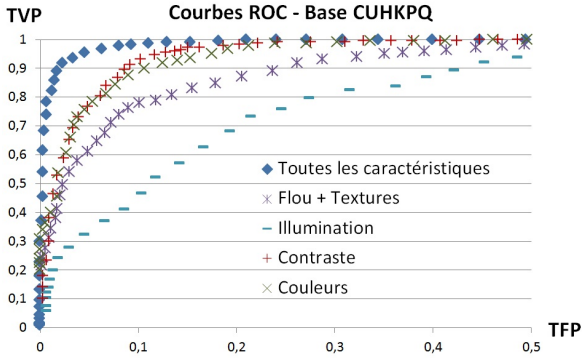


FIGURE 3.25 – Performances comparées de chaque catégorie de caractéristique.

TABLEAU 3.8 – Performances de classification binaire sur les 3 bases HFS, FAVA et PNF.

Base	HFS	FAVA	PNF
T_{BC} (%)	$79{,}9 \pm 0{,}9$	$66{,}8 \pm 1{,}3$	$63{,}3 \pm 0{,}3$

lisons que deux valeurs afin de décrire cet aspect, et que ces deux valeurs sont en outre très corrélées entre elles (moyennes des canaux L^* et V). De plus, ces valeurs ne représentent que des valeurs de pixels moyens sur une région donnée, et n'apportent aucune information sur les variations d'éclairage dans cette région. Intégrer des mesures d'illumination plus complexes est une possibilité d'extension de notre travail.

Ces observations montrent que nos régions et nos caractéristiques sont pertinentes : supprimer certaines d'entre elles diminue les performances de classification. En utilisant toutes nos caractéristiques et toutes les régions, nous obtenons un taux de bonne classification de 94% ainsi qu'une aire sous la courbe ROC de 0,99. Ces résultats suggèrent que nos caractéristiques et nos algorithmes d'apprentissage permettent en effet de distinguer deux catégories d'images aux scores de qualité esthétique extrêmes.

Bases de photos réalistes - HFS, FAVA et PNF

La base CUHKPQ ne contient que des photos de qualité esthétique extrêmes, tandis que les autres bases d'images contiennent un nombre important d'images moyennes. Afin d'effectuer de la classification binaire, pour chaque base d'images, les photos dont le score de qualité esthétique est inférieur au score médian sont associées à la classe "Mauvaise qualité esthétique", les autres sont associées à la classe "Bonne qualité esthétique". Nous cherchons alors à distinguer ces deux catégories. En respectant le protocole défini en 3.7, nous obtenons les résultats présentés dans le tableau 3.8.

Il ressort de ces expériences que les photos de la base HFS sont généralement correctement classées (autour de 80% de taux de bonne classification), tandis que les photos des bases FAVA et PNF sont plus souvent mal classées (entre 65 et 70% de taux de bonne classification). Cette

TABLEAU 3.9 – Performances de classification binaire sur les 3 bases HFS, FAVA et PNF, lorsque les photos de score intermédiaire ne sont pas prises en compte.

Base	HFS	FAVA	PNF
T_{BC} (%)	$84{,}7 \pm 1{,}6$	$74{,}3 \pm 1{,}3$	$68{,}4 \pm ,5$

différence s'explique par la présence d'un très grand nombre de photos dont le score est proche du score médian pour ces deux bases.

Influence de la suppression des photos moyennes

Nous découpons maintenant les bases d'images en 3 tiers égaux en effectif : les photos de très mauvaise qualité esthétique (le tiers des photos dont le score est le plus faible), les photos moyennes (le tiers des photos dont le score est proche du score médian), et les photos de très bonne qualité esthétique (le tiers restant, dont le score est le plus élevé). Nous reproduisons les expériences de classification binaire en ne tenant pas compte de la catégorie intermédiaire, et nous obtenons les résultats indiqués dans le tableau 3.9. Nous observons que les performances de classification sont plus élevées, car la majorité des photographies difficiles à classer sont celles dont le score de qualité esthétique est proche du score médian. Les taux de bonne classification obtenus pour la base PNF n'évoluent pas significativement car même en supprimant un tiers des photographies moyennes, il existe toujours un très grand nombre de photographies proches du score médian (voir figure 3.4). En supprimant les deux tiers des photographies les plus proches du score médian, les taux de bonne classification augmentent jusqu'à $76{,}4\% \pm 0{,}5$ pour la base PNF.

Toutefois, dans une situation réelle, nous avons également à prendre des décisions concernant les photographies de qualité esthétique moyenne. Une solution permettant de tenir compte de ces photos est la classification à 3 catégories.

3.7.2 Classification étendue à 3 catégories

Nous créons dans cette section une troisième catégorie de photographies. L'objectif est alors de correctement classer les photos, en évitant au maximum de confondre des photos correspondant aux catégories extrêmes. Par exemple, il est plus grave de confondre une photo de très bonne qualité esthétique avec une photo de très mauvaise qualité esthétique qu'avec une photographie moyenne.

Nous découpons ici également les bases d'images en 3 tiers égaux en effectif, correspondant respectivement aux photos de score très faible, moyen, et très élevé. Les algorithmes de classification sont ensuite appliqués, et les résultats obtenus sont indiqués dans le tableau 3.10. Ce tableau donne les différentes valeurs obtenues pour les mesures d'erreur CCE et MCE (voir équations 2.15 et 2.16) pour chacune des bases. Les valeurs de CCE indiquées sont normalisées par rapport au nombre d'images, ce qui signifie que $CCE(0)$ correspond au taux de

TABLEAU 3.10 – Performances de classification à 3 catégories sur les bases HFS, FAVA et PNF. Les scores sont donnés en pourcentage du nombre total d'image. Les valeurs de $CCE(2)$ et $CCE(-2)$ doivent être les plus faibles possibles.

Base	$CCE(-2)$	$CCE(-1)$	T_{BC}	$CCE(1)$	$CCE(2)$	MCE
HFS	2,1	16,3	$65{,}4 \pm 1{,}5$	13,8	1,7	44,7
FAVA	2,9	20,7	$49{,}0 \pm 0{,}8$	22,9	4,1	65,4
PNF	5,5	23,4	$44{,}1 \pm 0{,}3$	20,9	6,2	76,1

TABLEAU 3.11 – Performances de régression sur les bases HFS, FAVA et PNF, mesurées par les coefficients de corrélation de Pearson (R) et de Spearman (ρ).

Base	SVM		FAVA		PNF	
Coefficient	R	ρ	R	ρ	R	ρ
Résultat (%)	$73{,}8 \pm 1{,}3$	$74{,}2 \pm 1{,}4$	$49{,}0 \pm 1{,}3$	$49{,}1 \pm 1{,}4$	$42{,}9 \pm 0{,}6$	$42{,}1 \pm 0{,}6$

bonne classification. Les valeurs de $CCE(1)$ et $CCE(-1)$ correspondent au taux d'erreurs peu importantes (images moyennes confondues avec des images de bonne ou de mauvaise qualité esthétique), tandis que les valeurs de $CCE(2)$ et $CCE(-2)$ correspondent à des taux d'erreurs importantes (images de bonne qualité esthétique confondues avec des images de mauvaise qualité esthétique). Nous pouvons observer dans le tableau 3.10 que les valeurs de $CCE(2)$ et $CCE(-2)$ sont toujours significativement inférieures à celles de $CCE(1)$ et $CCE(-1)$, et également très faibles en valeur absolue. En effet, même pour la base PNF, seules environ 10% des images sont sur-évaluées ou sous-évaluées de 2 catégories ($CCE(2) + CCE(-2)$). Concernant la base HFS, nous observons des taux d'erreurs très faibles ($CCE(2) + CCE(-2) < 5\%$) : nos caractéristiques et nos algorithmes nous permettent de trier correctement les photos en 3 catégories, en évitant au maximum de confondre des images de très bonne et de très mauvaise catégories.

Remarquons enfin que les taux de bonne classification de la base PNF sont plutôt faibles ($44{,}1 \pm 0{,}3\%$). Ces taux sont significativement inférieurs à ceux observés pour la base HFS ($65{,}4 \pm 1{,}5\%$), mais toujours strictement supérieurs au hasard (33%). Comme nous l'avons déjà évoqué en 3.7.1, il n'est peut-être pas suffisant de découper la base d'images en 3 groupes d'effectifs égaux pour évaluer la base PNF. Afin d'affiner les résultats, il est alors possible d'effectuer de la régression.

3.7.3 Régression

Nous cherchons dans cette section à attribuer un score aux photographies. Les performances sont mesurées à l'aide des coefficients de corrélation de Pearson et de Spearman, qui mesurent la corrélation entre les scores de vérité terrain et ceux fournis par les algorithmes d'apprentissage. Les résultats obtenus sont indiqués dans le tableau 3.11. Ici également, nous observons que les prédictions concernant la base HFS sont plus précises que celles effectuées sur les bases FAVA et PNF.

Finalement, nous résumons les contributions apportées par nos travaux sur la figure 3.26. La figure 3.26 **a)** présente les résultats obtenus pour la base HFS avec les méthodes actuelles de l'état de l'art : les caractéristiques sont calculées sur la région du visage en entier, tandis que nous incluons également des informations provenant de régions plus fines (les yeux, la bouche). Le nuage de points obtenu pour la base HFS en intégrant ces informations est présenté sur la figure 3.26 **b)**. En intégrant les algorithmes de sélection de caractéristiques et de fusion des scores, nous obtenons le nuage de points **c)**, qui présente une corrélation entre les scores de vérité terrain et de prédiction significativement plus élevée que le nuage de points **a)**. Enfin, le nuage de points obtenus par régression sur la base FAVA est donné en figure 3.26 **d)**.

3.8 Comparaison avec l'état de l'art

Très peu de travaux ont été réalisés concernant l'estimation de la qualité esthétique des photos de visage. Certaines méthodes permettent toutefois d'évaluer des images dont les contraintes sont plus souples : photos contenant des visages, portraits de personnes représentées en entier, etc. Afin de comparer nos travaux à l'état de l'art, dans cette section nous incluons dans notre jeu de caractéristiques les informations sur la taille et la position du visage (voir section 3.3.2). Nous montrons ainsi au passage que nos travaux sur les photos de visage sont applicables à d'autres types de photos contenant des visages.

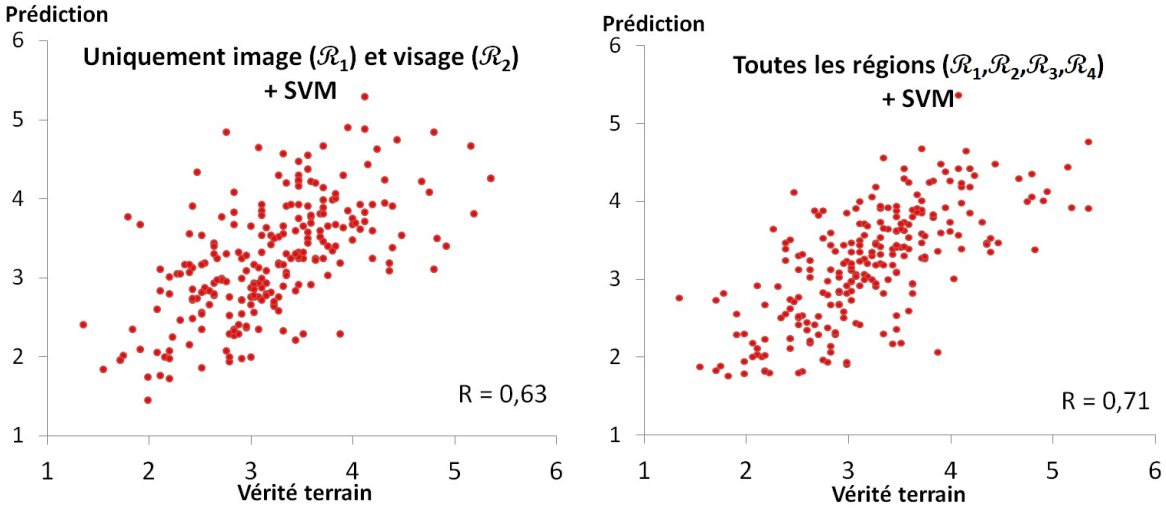
Pour toutes les comparaisons effectuées dans cette section, nous utilisons les 40 meilleurs couples de caractéristiques sélectionnés par l'algorithme Relief. Les algorithmes considérés pour l'apprentissage sont précisés pour chaque expérience.

3.8.1 Performances de classification

CUHKPQ

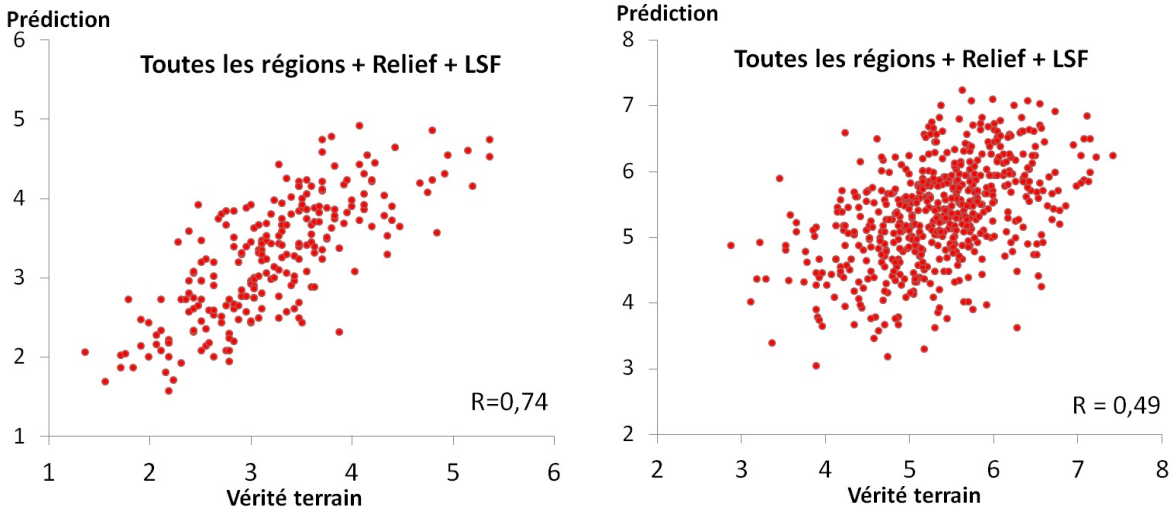
CUHKPQ est utilisé dans plusieurs travaux récents [Tang et al. 2013 ; Kim et Kim 2014] afin d'étudier la séparabilité d'images de très bonne et de très mauvaise qualité esthétique. Pour ce type de photos, Tang et al. et Kim et Kim obtiennent respectivement des aires sous la courbe ROC (AUC) de 0,974 et de 0,972. Ces deux travaux reposent sur un algorithme de segmentation de l'avant-plan et de l'arrière-plan, et sur le calcul de caractéristiques dans ces régions. Cette méthode est semblable à nos travaux, lorsque nous n'utilisons pas les informations contenues sur les yeux et la bouche. Sans ces informations, nous obtenons d'ailleurs des résultats très similaires ($AUC = 0{,}968$). En utilisant notre jeu de caractéristiques intégrant les informations issues des régions des yeux et de la bouche ainsi que l'algorithme d'apprentissage SVM, nous dépassons ces valeurs et obtenons une aire sous la courbe ROC de 0,990. Ceci correspond à un taux de bonne classification de 94% (contre environ 92% pour les travaux précédents).

FAVA



(a) Nuage de points obtenu par SVM pour HSF en utilisant uniquement les informations du visage et de l'image entière.

(b) Nuage de points obtenu par SVM pour HSF en utilisant uniquement toutes nos informations.



(c) Nuage de points obtenu par LSF pour HSF en utilisant uniquement les informations pertinentes.

(d) Nuage de points obtenu par LSF pour FAVA en utilisant uniquement les informations pertinentes.

FIGURE 3.26 – Différents nuages de points obtenus pour les bases HFS et FAVA dans différentes conditions. L'objectif est d'obtenir une prédiction proportionnelle à la vérité terrain ($R = 1$).

Nous considérons maintenant la base FAVA. Sur cette base également, les travaux précédents se limitent à l'étude des portraits sans les contraintes sur la position et la taille du visage définies dans l'introduction. Toutefois [Pogačnik et al. 2012] étudient des photos très proches des nôtres : le visage est suffisamment grand et proche du centre. Dans leurs travaux, Pogačnik et al. considèrent au total 1048 images issues du site DPChallenge. Les images dont le score de vérité terrain est moyen (entre 4,5 et 6,5 sur une échelle de 1 à 10) sont exclues. Nous faisons de même sur la base FAVA ; les images dont le score de vérité terrain est inférieur à 4,5 sont regroupées dans une catégorie, celles au-dessus de 6,5 dans l'autre catégorie. Pogačnik et al. obtiennent un taux de bonne classification de 73,4% en utilisant 71 caractéristiques (nous en utilisons 60), et améliorent les performances jusqu'à 74,8% en conservant les 41 meilleures caractéristiques proposées par l'algorithme Relief. En utilisant nos caractéristiques, les algorithmes SVM et Relief, nous obtenons un taux de bonne classification de l'ordre de 81% (soit une augmentation de 6%).

AVA

Récemment, les travaux de [Redi et al. 2015] proposent une évaluation des portraits contenus dans la base AVA. Au total, 10141 photos contenant des visages sont utilisées, sans contrainte particulière sur la taille et la position des visages. Dans ces travaux, un grand nombre de caractéristiques décrivant les visages sont extraites : attributs correspondant à la personne représentée (estimation de l'âge, du sexe, présence de sourire), statistiques calculées dans différentes régions (comme celles décrites dans ce document), etc. Un point particulièrement intéressant est l'utilisation d'informations dans les régions définies par les yeux et la bouche : netteté, illumination, teinte. Cette idée est justement celle évoquée dans ce chapitre, et les travaux de [Redi et al. 2015] présentent les mêmes conclusions quant aux caractéristiques les plus pertinentes pour l'estimation de la qualité esthétique des photos de visage : la netteté calculée dans les régions des yeux et de la bouche est la caractéristique la mieux corrélée aux scores de vérité terrain. Ceci confirme notre intuition et surtout la pertinence des méthodes et résultats présentés jusqu'ici. Aussi, en séparant les images en deux catégories (scores faibles, scores élevés), [Redi et al. 2015] obtiennent un taux de bonne classification de $64{,}2 \pm 1{,}8\%$, tandis qu'en utilisant le même algorithme de classification (SVM) sans sélection de caractéristiques et sur exactement les mêmes images, nous obtenons $64{,}8 \pm 0{,}3\%$. Nous améliorons les performances de classification à $65.6 \pm 0{,}2\%$ en fusionnant les résultats de nos 4 algorithmes après sélection des caractéristiques.

Flickr

[Li et al. 2010a] proposent des expériences de classification et de régression sur la base de 500 photos contenant des visages obtenues sur le site [*Flickr*]. Pour la classification, leur objectif est de distinguer 5 catégories d'images : photos de très mauvaise, mauvaise, moyenne, bonne et très bonne qualité esthétique. Pour cela, 5 catégories de 100 photos sont constituées en fonction des scores de vérité terrain : les 100 photos de plus mauvaise qualité esthétique sont regroupées dans une première catégorie, puis les 100 suivantes dans la seconde, etc. Leurs

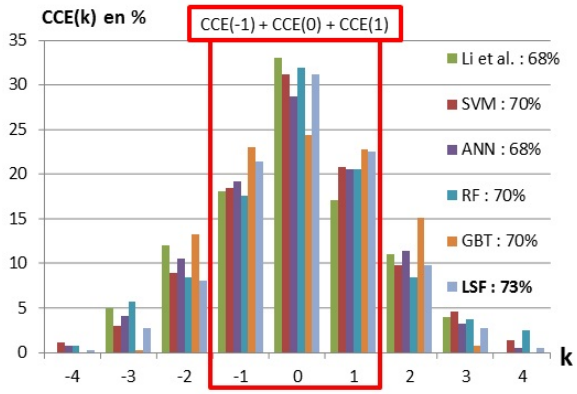


FIGURE 3.27 – Valeurs du CCE rapportées en pourcentage du nombre d'images total, pour la classification à 5 catégories sur la base Flickr. Les différents algorithmes d'apprentissage sont comparés aux travaux de Li et al.

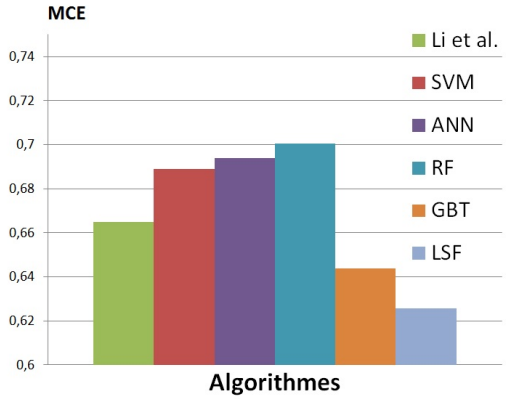


FIGURE 3.28 – Valeurs du MCE, pour la classification à 5 catégories sur la base Flickr. Les différents algorithmes d'apprentissage sont comparés aux travaux de Li et al.

résultats de classification peuvent être évalués grâce aux valeurs définies par le CCE et le MCE dans les équations 2.15 et 2.16. Nous effectuons la même expérience sur ces images avec notre jeu de caractéristiques, pour lesquels les résultats sont donnés sur les figures 3.27 et 3.28. Nous voyons sur ces figures tout l'intérêt de la fusion des scores. En effet, si SVM et RF permettent d'obtenir le meilleur taux de bonne classification (soit une valeur de $CCE(0)$ élevée), l'algorithme GBT permet de diminuer significativement le MCE (peu d'erreurs importantes). [Li et al. 2010a] évaluent les performances globales de la classification en comptant le nombre d'images pour lesquelles l'erreur de classification est au plus d'une classe (voir le cadre rouge de la figure 3.27), ce qui correspond concrètement à la somme $CCE(-1)+CCE(0)+CCE(1)$. En reprenant ce critère, nous obtenons à peu près les mêmes performances lorsque nous utilisons nos algorithmes séparément : entre 68 et 70% pour nos algorithmes et environ 68% pour Li et al. La fusion des prédictions permet d'améliorer significativement les performances globales : $CCE(-1) + CCE(0) + CCE(1) = 73\%$. Ceci nous montre non seulement que nous pouvons évaluer correctement tout type de photos contenant des visages à l'aide de notre méthode (la composition des photos utilisées ici est très libre), mais également que nous pouvons obtenir des résultats supérieurs à ceux existants.

Notons que pour l'évaluation de photos contenant plusieurs personnes, en plus des aspects compositionnels (taille, position des visages) également utilisés pour l'évaluation des portraits, il serait intéressant d'intégrer dans nos modèles les aspects relationnels entre ces visages : distance entre les visages, orientation commune des visages, etc. Nous pensons que l'ajout de tels critères, déjà utilisés par [Li et al. 2010a], permettrait d'améliorer encore les résultats.

Parmi les 500 photos de [Li et al. 2010a], environ 145 sont des portraits : une seule personne y est représentée. Les travaux de [Khan et Vogel 2012], sur l'étude particulière de la qualité esthétique des portraits, sont évalués sur ces 145 images. Ces images sont séparées en deux groupes, correspondant aux images de bonne et de mauvaise qualité esthétique. Leurs

TABLEAU 3.12 – Comparaison des performances (T_{BC} en %) obtenues sur les différentes bases de données entre les travaux de l'état de l'art et notre méthode, en utilisant un seul algorithme de classification (SVM) puis en utilisant l'algorithme Relief ainsi que notre algorithme de fusion des scores (LSF). Les meilleurs scores sont indiqués en gras.

Base	Kim et al.	Poga. et al.	Redi et al.	Li et al.	Khan et al.	SVM	LSF
CUHKPQ	92	/	/	/	/	94	**95**
FAVA	/	75	/	/	/	**81**	**81**
AVA	/	/	64	/	/	65	**66**
Flickr	/	/	/	68	/	70	**73**
FlickrP	/	/	**74**	/	63	67	70
HFS	/	/	/	/	/	79	**80**
PNF	/	/	/	/	/	64	**65**

performances de classification sont de l'ordre de 63% tandis que nous atteignons 67% avec le même algorithme de classification (SVM). Il est même possible d'atteindre 69,5% de bonne classification en fusionnant les prédictions obtenues par les différents algorithmes. Toutefois, dans ce cas précis, nos performances sont en-dessous du meilleur résultat de l'état de l'art (74%), proposé par [Redi et al. 2015]. Ceci peut s'expliquer par le fait que certaines informations sur les visages ne sont pas prises en compte dans notre modèle (la présence de sourire par exemple). Notons aussi qu'à cause du très faible nombre d'images dans la base (145), les résultats obtenus sont très variables : l'écart-type des performances de nos algorithmes sur FlickrP est de l'ordre de 3%.

Bilan des performances de classification

Les performances de classification sont résumées dans le tableau 3.12. Nous y voyons que dans la plupart des cas, nous égalons ou dépassons les performances de l'état de l'art. Si les contraintes (taille et position du visage) sur les photos sont élevées, alors la pertinence et les performances des caractéristiques et des régions considérées dans ce travail sont également élevées.

Remarque : Pour obtenir les résultats concernant les bases AVA et PNF, nous avons dû modifier les paramètres correspondant aux critères d'arrêt des algorithmes SVM et ANN. En effet, nous avons observé qu'avec les conditions définies en 2.4, certains des algorithmes d'apprentissage n'ont pas le temps de converger : le nombre d'images étant plus conséquent (respectivement plus de 10000 et plus de 28000 photos, contre moins de 1000 pour les autres bases). En remplaçant le nombre maximal d'itérations autorisé pour la convergence de ces algorithmes par une valeur proportionnelle au nombre de photos de la base d'apprentissage, les performances sont significativement améliorées. Il est également possible d'augmenter le nombre d'arbres utilisés pour les forêts aléatoires et boostées, ce qui augmente légèrement les performances de prédiction. Cette remarque vaut également pour les résultats de régression.

3.8.2 Performances de régression

Flickr

Si peu de travaux ont été effectués sur l'estimation de la qualité esthétique de photos de visage, il existe encore moins de travaux dans lesquels le problème de la régression a été abordé. En effet, celui-ci est plus complexe, car les modèles doivent être capables d'évaluer précisément une image plutôt que de deviner une catégorie. Un premier modèle de régression concernant les photos contenant des visages est celui proposé par [Li et al. 2010a]. Ce modèle est obtenu par apprentissage automatique (SVM linéaire) à partir des mêmes caractéristiques que celles utilisées dans leur modèle de classification : mesures de composition de la photo (taille, position des visages), de leurs relations (distances entre les visages, expressions et orientations des visages), et d'autres caractéristiques telles que les couleurs et le contraste de l'image. Les performances sont évaluées à l'aide de l'erreur quadratique moyenne (voir équation 2.18). Sur les 500 images de la base Flickr, l'erreur quadratique moyenne obtenue pour Li et al. est de 2,28. Nous obtenons des valeurs d'erreur quadratique moyenne de 2,12 avec nos caractéristiques et l'algorithme LSF. Ce dernier résultat est similaire aux résultats obtenus par les travaux de [Xue et al. 2013] ($EQM = 2{,}11$), dans lesquels des informations plus complexes à calculer sont prises en compte : présence de sourire, orientation du visage. Pour obtenir leurs résultats, Li et al. et Xue et al. ont besoin de calculer des informations complexes et sensibles aux performances d'algorithmes d'analyse faciale, dont nous parlons au chapitre 4. Ceci rend notre méthode plus facilement reproductible (celle-ci dépend uniquement de la détection de nos régions d'intérêt), pour des performances tout à fait similaires. Nous indiquons également dans le tableau 3.13 les valeurs que nous obtenons pour les coefficients de corrélation de Pearson et de Spearman pour cette base d'images.

AVA

De par la présence d'un grand nombre d'images évaluées selon leur qualité esthétique, des travaux récents sur l'estimation de la qualité esthétique de photos utilisent la base AVA. [Redi et al. 2015] observent $\rho = 0{,}40$ en utilisant des caractéristiques similaires aux nôtres, plus des caractéristiques nécessitant l'utilisation d'algorithmes d'analyse faciale. Sur les mêmes 10141 images que celles étudiées par [Redi et al. 2015], nous obtenons $\rho = 0{,}41$. Rappelons que ces images ne sont pas contraintes sur la taille et la position du visage, et cela montre encore une fois que nos caractéristiques sont adaptées même sur des images à la composition moins stricte. En ajoutant les informations sur la taille et la position des visages dans nos modèles, nous obtenons $\rho = 0{,}42$. En appliquant l'algorithme Relief ainsi que la fusion des prédictions, nous arrivons à $\rho = 0{,}45$: les différents outils que nous avons introduits dans le chapitre 2 permettent d'améliorer significativement les performances de régression.

Si nous nous restreignons à l'ensemble des photos de la base FAVA, dans laquelle les images respectent les contraintes définies en introduction, nous obtenons $\rho = 0{,}50$. Ce résultat est significativement supérieur à celui obtenu dans le cas de photos de visage non contraintes et ceci montre que nos caractéristiques sont particulièrement adaptées aux photographies de

visage.

Bilan des performances de régression

Nous indiquons les performances obtenues par régression sur toutes nos bases de photos dans le tableau 3.13. Les très bons résultats de corrélation sur la base HFS (R et $\rho > 0{,}7$) prouvent que les caractéristiques que nous proposons permettent d'évaluer très correctement des photographies d'amateurs, ce qui est notre objectif principal. Il est donc possible d'appliquer notre méthode directement dans un logiciel triant automatiquement les photos selon leur qualité esthétique. Nous proposons un exemple d'application dans le dernier chapitre de ce document. A l'inverse, les autres bases de photos (notamment FAVA, PNF) contiennent surtout des photos prises par des photographes professionnels, pour lesquels il est plus difficile de quantifier la qualité esthétique à l'aide de nos algorithmes et jeux de caractéristiques (R et $\rho < 0{,}5$).

TABLEAU 3.13 – Comparaison des performances (EQM, R ou ρ) obtenues sur les différentes bases de données entre les travaux de l'état de l'art et notre méthode, en utilisant un seul algorithme de régression (SVM) puis en utilisant l'algorithme Relief ainsi que notre algorithme de fusion des scores (LSF). Les meilleurs scores sont indiqués en gras.

Base	Li et al.	Xue et al.	Redi et al.	SVM			LSF		
	EQM	EQM	ρ	EQM	R	ρ	EQM	R	ρ
Flickr	2,28	**2,11**	/	2.15	0,47	0,49	**2,12**	**0,48**	**0,50**
AVA	/	/	0,40	0,58	0,42	0,42	**0,55**	**0,44**	**0,45**
FAVA	/	/	/	0,66	0,43	0,44	**0,59**	**0,49**	**0,50**
HFS	/	/	/	**0,28**	0,73	**0,74**	0,29	**0,74**	**0,74**
PNF	/	/	/	0,46	0,42	0,41	**0,45**	**0,43**	**0,42**

3.9 Application à la recherche et la sélection de photographies de visage

3.9.1 Protocole proposé

Dans cette partie, nous cherchons à retrouver les photos de visage à la qualité esthétique la plus élevée, dans une base d'images ne contenant pas uniquement des photos de visage, mais des photos de toute catégorie. Nous procédons selon le protocole suivant.

Dans un premier temps, nous effectuons une régression sur une base annotée de photos de visage. Les descripteurs considérés pour l'apprentissage sont les 40 couples (Caractéristique, Région) dont la valeur renvoyée par l'algorithme Relief est la plus élevée. Nous choisissons d'effectuer la régression à partir des photos de la base HFS, qui respectent nos contraintes sur la position et la taille du visage, tout en étant annotées par des humains dans un environnement

contrôlé.

Ensuite, pour chaque photographie de la base Flickr (donc non utilisée lors de l'apprentissage), nous appliquons les étapes suivantes :

1. Détection du visage dans la photo. Si aucun visage n'est détecté, la photo est ignorée. Si plusieurs visages sont détectés, dans cette expérience, seul le visage le plus grand est considéré.
2. Extraction du visage de la photo. La photo initiale est rognée de manière à ce que le visage détecté respectent nos contraintes de taille et de position dans l'image.
3. Évaluation de la photo de visage obtenue. Le modèle de régression appris sur la base HFS est utilisé afin d'évaluer la nouvelle photo.

Enfin, seules les photos dont l'évaluation de la qualité esthétique est élevée sont sélectionnées. Ce protocole nous permet par exemple de retrouver, dans une base d'images conséquente, toutes les photographies de visage de bonne qualité esthétique. Il est possible de coupler cet outil avec des méthodes de reconnaissance faciale afin de proposer les photos de qualité esthétique élevée d'une personne donnée.

3.9.2 Photos sélectionnées

Nous présentons les photos de visage sélectionnées à l'aide de ce protocole sur la figure 3.29. Les prédictions sont effectuées sur les visages détectés par l'algorithme, et cette figure présente uniquement les photos de la base pour lesquelles les scores de prédiction des visages extraits sont élevés. Nous y voyons que les couleurs sont généralement vives, le visage est net et contrasté par rapport à l'arrière-plan.

A l'inverse, nous observons sur la figure 3.30 les visages ayant les scores de qualité esthétique les plus faibles selon l'algorithme. Ces photos de visage présentent plutôt des couleurs ternes, le visage peut être flou et le contraste à l'intérieur du visage est très faible.

Les performances sont évaluées visuellement car nous n'avons pas accès aux scores de qualité esthétique des visages extraits : seuls les scores des photos entières sont disponibles.

3.9.3 Temps de calcul

L'application que nous venons de décrire a pour objectif d'aider à la sélection de photographies, or si le temps de calcul est trop contraignant, l'intérêt pratique de notre méthode est limité. Il serait également intéressant de faire fonctionner ce type d'algorithme en temps réel afin de pouvoir intégrer ce type d'algorithme dans un appareil photo ou pour traiter des vidéos.

Nous avons déjà vu que l'algorithme de détection des attributs du visage nous permet de traiter entre 5 et 30 images par seconde, ce qui constitue une première limite. Nous exposons dans le tableau 3.14 les temps de calcul concernant la détection des attributs faciaux ainsi que

FIGURE 3.29 – Photographies de bonne qualité esthétique sélectionnées par l'algorithme. Les photos initiales sont fournies à l'algorithme, qui détecte les visages, les extrait de la photo, puis les évalue et les trie.

pour l'extraction des caractéristiques, pour différentes tailles d'images. Il ressort que le temps de calcul des caractéristiques est 10 fois plus élevé que le temps de détection. Cela s'explique par le calcul du canal sombre, qui est un filtre minimal nécessitant un grand nombre d'opérations. Nous avons constaté que le fait de ne pas calculer le canal sombre dans l'image ou le visage entier ne diminue pas les performances, et accélère très grandement le temps de calcul. Cette optimisation est visible à la troisième ligne du tableau.

Finalement, même pour des images de très grande taille, il est possible de traiter au moins 3 photos par seconde. Pour des images très petites (par exemple lorsque le visage est extrait d'une photo plus grande), il est possible de traiter environ 25 images par seconde. Il est toujours possible d'accélérer la vitesse d'exécution en sous-échantillonnant les images au

TABLEAU 3.14 – Temps de calcul moyen par image, pour différentes tailles d'images. Les temps indiqués correspondent au temps de détection des attributs (Détection), d'extraction des caractéristiques (Extraction) et d'extraction sans le calcul du canal sombre dans les régions 3 et 4 uniquement (Extraction opt.).

Taille de l'image	240 × 180 pixels	640 × 480 pixels	1600 × 1200 pixels
Détection	40ms	120ms	210ms
Extraction	20ms	460ms	2150ms
Extraction opt.	6ms	30ms	110ms

FIGURE 3.30 – Photographies reconnues par l'algorithme comme étant de mauvaise qualité esthétique. Les photos initiales sont fournies à l'algorithme, qui détecte les visages, les extrait de la photo, puis les évalue et les trie.

préalable (au moins pour la phase de détection), au risque de ne pas pouvoir détecter de trop petits visages dans les photos. Ne pas calculer certaines informations (le canal sombre par exemple) sur l'image entière permet également de réduire le temps de calcul sans diminuer significativement les performances de prédiction. Enfin, nous n'avons pas inclus les temps de prédiction dans cette analyse, car ceux-ci sont largement négligeables devant les temps de détection du visage et d'extraction des caractéristiques.

3.10 Conclusion

3.10.1 Avantages de la méthode

Dans ce chapitre, nous avons présenté une méthode de prédiction de la qualité esthétique adapté aux photos de visage. Nous avons montré qu'il est tout à fait possible d'adapter nos travaux à tout type de photos contenant des visages, et d'obtenir des performances similaires ou supérieures à celles de l'état de l'art. De plus, nous pouvons très facilement renforcer la précision de nos prédictions en intégrant des informations liées à la composition de la photo : taille et position de chaque visage, relations entre les visages, estimation des expressions faciales (cf. les travaux de [Li et al. 2010a ; Xue et al. 2013 ; Redi et al. 2015]).

Un autre avantage lié à l'utilisation de nos caractéristiques est leur robustesse par rapport à la taille des images étudiées. En effet, le fait de sous-échantillonner une image ne modifiera que très peu nos mesures, basées essentiellement sur des moyennes ou des écarts-types dont les valeurs ne changent que très peu lorsque l'image est sous-échantillonnée. Cet avantage peut être exploité pour analyser très rapidement un grand nombre d'images, comme nous l'avons montré en 3.9.

Nous avons également montré que le jeu de caractéristiques proposé est pertinent quelle que soit la base d'images ou le problème considéré. Nous avons constaté que si certaines caractéristiques sont plus pertinentes que d'autres pour un problème particulier, celles-ci peuvent être mises en avant à l'aide de l'algorithme Relief. Ainsi, nous pouvons construire des modèles plus précis et moins complexes car basés sur des jeux de caractéristiques plus restreints, décrivant plus précisément la qualité esthétique des images. Cette étape permet aussi d'analyser les éléments discriminants pour la qualité esthétique, et donc d'obtenir des informations pertinentes sur ce qui fait une belle photo.

Enfin, le fait de fusionner les prédictions de plusieurs algorithmes rend le modèle moins dépendant d'un algorithme particulier et de caractéristiques particulières. En effet, selon les données, différents algorithmes présentent des performances différentes, mais la fusion de leur prédiction produit une prédiction globale améliorée. Tout comme l'utilisation de l'algorithme Relief, la fusion des algorithmes permet d'augmenter la robustesse de nos modèles aux changements de type d'image ou de problèmes considérés.

3.10.2 Limites de la méthode

Les modèles que nous proposons sont efficaces car nous accédons à de l'information très précise contenue dans des régions particulières : les yeux, la bouche. Un point crucial de notre algorithme correspond ainsi à l'évaluation correcte de la position de ces éléments dans le visage. Si l'algorithme de Viola-Jones que nous utilisons permet d'obtenir de bonnes performances de détection, nous obtenons un taux de détections manquantes ou erronées non négligeable. Or nous ne pouvons évaluer correctement nos images sans l'information correspondant à ces régions. Plusieurs solutions peuvent être envisagées :

- Évaluer l'image en ne tenant compte que des informations obtenues sur toute l'image, ce qui en pratique signifie compléter les valeurs manquantes dans le vecteur décrivant l'image (voir au chapitre 2, en 2.2.1).
- Proposer une région "par défaut" (par exemple le tiers central de l'image serait la région du visage) dans laquelle les caractéristiques sont calculées lorsque le visage ou l'un de ses attributs n'est pas détecté.
- Simplement ignorer les photos ne contenant pas de visage. En effet, dans le cadre d'une application réelle, il est tout à fait possible de fournir en entrée des images ne contenant pas de visage.

Comme nous l'avons évoqué, le temps de calcul des caractéristiques est suffisamment faible. Toutefois, l'introduction de plusieurs algorithmes d'apprentissage ralentit d'autant la vitesse d'apprentissage ; dans notre exemple 4 apprentissages distincts sont nécessaires pour la construction de chaque modèle. De même, la prédiction globale nécessite le calcul des prédictions de chaque algorithme. Dans le cadre d'une application réelle, l'apprentissage n'est effectué qu'une fois au départ, tandis que le temps associé à la prédiction est très largement inférieur au temps de calcul des caractéristiques et ne pose donc pas de problème pour le temps de calcul global.

Enfin, notre méthode produit des performances significativement différentes sur des bases de données distinctes. En effet les photos de visage contenues par exemple dans les bases HFS et PNF contiennent respectivement des photos d'amateur, de qualité esthétique généralement moyenne ou basse, et des photos de photographes professionnels dont la qualité esthétique est généralement bien plus élevée. Ainsi, nous obtenons de très bonnes performances sur les bases d'images constituées de photographies d'amateurs (HFS, corrélation $R > 0{,}7$), tandis que la précision de nos prédictions est plus faibles lorsque les images considérées sont des photographies de professionnels (FAVA, PNF, corrélation $R < 0{,}5$).

3.10.3 Bilan

Nous avons décrit dans ce chapitre une méthode simple (peu de caractéristiques employées, extraction directe sur les pixels de l'image), facile à mettre en œuvre (utilisation d'algorithmes classiques tels que Viola-Jones pour la détection, SVM pour la classification), et produisant des résultats dépassant ceux de l'état de l'art. De plus les informations que nous obtenons sur les images permettent d'analyser les différents éléments déterminant la qualité esthétique

d'une photo de visage.

Nous avons également montré qu'une part très importante de l'information de qualité esthétique des photos de visage est contenue dans la région des yeux. Ces derniers sont ainsi des éléments essentiels de la photo, et leur analyse permet à elle seule d'établir des modèles pertinents. Cet aspect est l'apport principal de notre méthode par rapport aux travaux précédents, et semble très prometteur dans le sens où le calcul d'informations dans cette région est très rapide et donc utilisable en temps réel, notamment lorsque notre outil est connecté à un logiciel de suivi du visage.

Toutes les étapes de la méthode sont implémentées en C++ et directement utilisables sur des photographies quelconques contenant au moins un visage.

Chapitre 4

Estimation de l'impression véhiculée par une photo de visage : cas de la compétence et de la sympathie

Sommaire

4.1 Introduction . . . 129
4.2 Bases d'images annotées . . . 131
4.2.1 Images synthétiques . . . 131
4.2.2 Images réelles . . . 132
4.3 Extraction des caractéristiques . . . 135
4.3.1 Filtres de Gabor . . . 135
4.3.2 Position de points de repère sur le visage . . . 136
4.3.3 Attributs de haut niveau . . . 138
4.4 Étude de l'influence des caractéristiques . . . 141
4.4.1 Comparaison des 3 lots de caractéristiques . . . 142
4.4.2 Caractéristiques discriminantes . . . 143
4.4.3 Comparaison des outils d'extraction des attributs . . . 144
4.5 Estimation des impressions de compétence et de sympathie . . . 145
4.5.1 Classification binaire . . . 145
4.5.2 Régression . . . 146
4.6 Comparaison avec l'état de l'art . . . 147
4.6.1 Performances sur les visages synthétiques . . . 147
4.6.2 Performances sur les visages réels . . . 148
4.7 Fusion des modèles de qualité esthétique et de sympathie . . . 149
4.7.1 Méthode de fusion . . . 149
4.7.2 Exemples de sélection de photos pour une personne donnée . . . 150
4.8 Conclusion . . . 150

4.1 Introduction

Objectif

Dans ce chapitre nous nous intéressons essentiellement aux impressions de compétence et de sympathie véhiculées par une photo de visage. L'objectif est de pouvoir sélectionner automatiquement les photos dans lesquelles une personne donnée dégage une impression particulière.

Pour atteindre ce but, nous avons choisi une approche basée sur l'utilisation d'attributs de haut niveau. Ces attributs sont détaillés en 4.3.3 et représentent essentiellement des informations relatives au visage. Ces informations peuvent être des informations sur le visage (forme générale du visage, longueur des cheveux), les expressions faciales (présence de sourire, émotions, ouverture des yeux) ainsi que des caractéristiques non permanentes (lunettes, moustache, barbe, maquillage, etc.). Nous appliquons ensuite les outils proposés dans le chapitre 2 afin de proposer des modèles pertinents d'évaluation des impressions de compétence et de sympathie dégagées par une photo de visage.

Jusqu'à présent, les modèles de l'état de l'art reposent uniquement sur les positions de points de repères particuliers dans le visage ou sur des descripteurs génériques tels que les histogrammes de gradients orientés [Rojas et al. 2010 ; Rojas et al. 2011] qui permettent de mettre en évidence les contours des différents attributs du visage (des yeux et des sourcils, du nez, de la bouche) afin d'associer ces contours à une expression faciale particulière. Les expressions du visage sont en effet largement associées aux impressions véhiculées : un visage souriant exprimant de la joie est considéré comme sympathique, tandis qu'un visage colérique dégage un sentiment d'antipathie ou de menace. Dans notre travail, nous montrons que si ces éléments peuvent être pris en compte afin d'affiner les modèles, les différents attributs de haut niveau que nous considérons sont bien plus performants. Ces attributs sont par ailleurs déjà utilisés dans le cadre de la reconnaissance faciale, comme le suggèrent les travaux de [Kumar et al. 2009 ; Dantcheva et al. 2011].

D'autres travaux cherchent également à évaluer les impressions dégagées par une photo de visage. Par exemple, [Mazza et al. 2015] cherchent à associer chaque photo de visage à une de ces 3 applications particulières : réseau professionnel, partage entre amis, site de rencontre. Cet objectif est très proche du nôtre. L'approche de Mazza et al. est similaire à celle présentée dans ce document, dans le sens où des attributs de haut niveau d'interprétation sont également utilisés. La différence réside essentiellement dans le fait que les attributs de Mazza et al. sont obtenus à l'aide d'annotations humaines, tandis que nous extrayons ces informations automatiquement (quitte à introduire des erreurs de mesures, les algorithmes étant moins fiables que les humains).

Plan du chapitre

La structure de ce chapitre est la suivante. Tout d'abord, nous décrivons en 4.2 les bases d'images que nous utilisons pour valider nos modèles. Ces bases peuvent être classées en deux catégories. La première comprend des images synthétiques, générées artificiellement par l'outil [*FaceGen*], ne possédant ni arrière-plan, ni accessoires particuliers (lunettes, maquillage), ni cheveux. La seconde catégorie correspond à des images réelles, respectant les contraintes sur la taille et la position du visage définies en introduction de ce document. Nous présentons

ensuite en 4.3 les différentes caractéristiques que nous extrayons sur les images. Celles-ci correspondent à des informations de haut niveau extraites de l'image à l'aide d'outils entièrement automatiques. A des fins de comparaisons avec l'état de l'art [Rojas et al. 2010 ; Rojas et al. 2011], nous calculons également des informations de bas niveau (filtres de Gabor, positions de points de repère dans le visage). Nous étudions les caractéristiques les plus discriminantes en 4.4, puis nous appliquons nos algorithmes afin d'estimer les impressions de compétence et de sympathie véhiculées par les photos en 4.5. Nous comparons nos résultats à l'état de l'art en 4.6. Enfin, nous présentons une application possible de ce travail à la sélection automatique de photographies en section 4.7.

Les différentes contributions de nos travaux pour l'évaluation des impressions véhiculées par une photo sont les suivantes :

— Création d'une base de 140 photos, évaluées selon les impressions de compétence et de sympathie dégagées par les photos.
— Définition et l'utilisation d'attributs de haut niveau afin de créer des modèles d'évaluation de traits de caractères suggérés par une photo.
— Amélioration significative des performances de l'état de l'art à l'aide des attributs de haut niveau, en particulier concernant l'évaluation de la sympathie.
— Démonstration de la faisabilité du tri automatique de photographies en fonction de l'impression de sympathie véhiculée par une photo de visage, pour une personne donnée.

4.2 Bases d'images annotées

4.2.1 Images synthétiques

Une première base de 300 visages générés aléatoirement par [*FaceGen*] est décrite dans les travaux de [Todorov et Oosterhof 2011]. Ces visages synthétiques sont annotés par plus de 20 personnes sur une échelle de 1 à 10, selon différents traits de caractère, parmi lesquels figurent la compétence et la sympathie. Les moyennes des évaluations des 20 personnes constituent la vérité terrain que nos modèles cherchent à atteindre. Différents visages générés ainsi que leurs évaluations sont donnés sur la figure 4.1.

A partir de ces évaluations, [Todorov et Oosterhof 2011] proposent un premier modèle d'évaluation des impressions de compétence et de sympathie à partir de la structure du visage (dans [*FaceGen*], le visage est représenté par une grille de 2043 points de repère) ainsi que d'informations de réflectance liées aux textures et aux couleurs du visage (représentées par les 246×246 pixels du visage). Pour chaque trait, le modèle d'évaluation est obtenu par une régression linéaire sur les 300 visages annotés, après avoir réduit la dimension des caractéristiques par analyse en composantes principales (ACP) à 100 valeurs (50 pour la structure, 50 pour la réflectance).

Le modèle généré permet ensuite de procéder à l'étape de création de visages suivant ces modèles. Ainsi, 25 visages différents sont créés, et les caractéristiques de ces visages sont ensuite

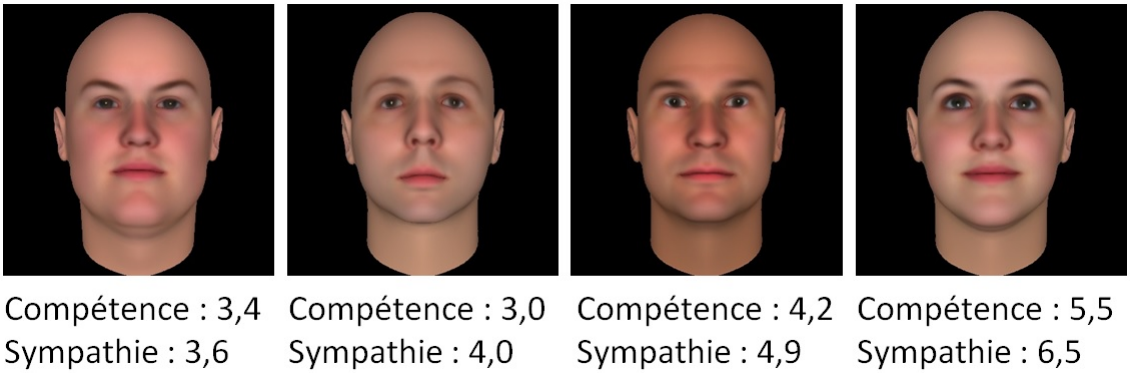


FIGURE 4.1 – Exemples de 4 photos de visage synthétiques. Les visages peuvent paraître compétents et sympathiques, sympathiques mais peu compétents, antipathiques et incompétents, etc. L'échelle des scores va de 1 à 10.

exagérées (déformation du visage, changement de la couleur de la peau...) afin d'augmenter ou d'atténuer le trait de caractère considéré, en fonction des modèles de forme et de réflectance appris par la régression. Chaque visage est alors modifié sur 7 niveaux d'expression de ce trait, par exemple de très antipathique à très sympathique pour le trait "sympathie". Nous obtenons ainsi 2 bases (une pour la sympathie, une pour la compétence) de 7×25 visages.

Les travaux de [Todorov et al. 2013] montrent que les deux bases ainsi créées s'approchent des évaluations humaines. En effet, en faisant évaluer les visages présents dans les deux bases de 175 visages par des participants à une expérience en laboratoire, de très fortes corrélations entre les jugements humains et les niveaux d'expression des caractères sont observées. Ce type de modèle est cependant difficile à généraliser sur des images réelles à cause de l'absence d'un grand nombre d'éléments extérieurs (cheveux, vêtements, arrière-plan, visages masculins uniquement).

Les bases d'images synthétiques sont utilisées avant tout pour déterminer les caractéristiques discriminantes pour l'estimation du degré de sympathie et de compétence suggéré par une photo de visage.

4.2.2 Images réelles

Base de visages existante (Karolinska)

A notre connaissance, la seule base de photos de visage disponible et annotée selon différents traits de caractère est celle présentée dans les travaux de [Todorov et al. 2008]. Celle-ci contient 66 visages d'hommes et de femmes. Il peut être intéressant de tester nos méthodes sur cette base, toutefois de la même manière que pour les visages synthétiques, il est difficile de généraliser un modèle à partir de celle-ci : même fond pour toutes les photos, même pose, même éclairage, pas d'accessoires ni d'expressions particulières. Le faible nombre d'images dans la base réduit également les chances d'obtenir un modèle d'évaluation précis. Des exemples de

visage et leurs évaluations sont données sur la figure 4.2. Notons que ces photos ne sont pas évaluées selon les caractéristiques de compétence et de sympathie, mais selon les 2 axes que sont la confiance et la dominance. Toutefois, à l'aide d'une analyse en composantes principales, les travaux de [Todorov et al. 2008] montrent que les autres traits de caractère (dont font partie la compétence et la sympathie) peuvent s'exprimer en fonction d'une combinaison linéaire de ces deux traits. Cette première base de photos est utilisée dans le but de comparer nos travaux à l'état de l'art.

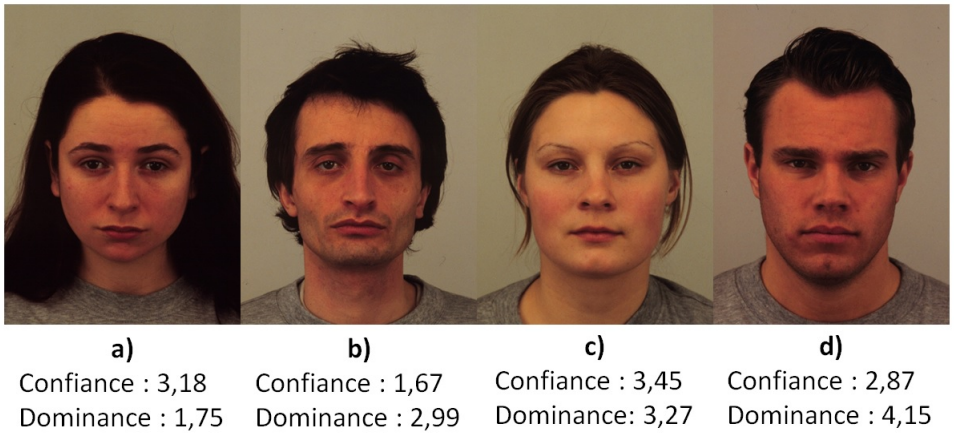


FIGURE 4.2 – 4 photos de visage de la base Karolinska, évaluées selon la confiance et la dominance. Le score moyen est de 3 pour chacune des catégories, l'échelle des scores va de 1 à 5.

Base de visages constituée dans le cadre de ce travail (HFS_{CS})

Afin de pallier le manque de bases annotées existantes, nous avons mené une expérience dans laquelle des participants ont évalué 140 photos de visage. Ces photos sont un sous-ensemble de la base HFS et correspondent à 7 photos de 20 personnes différentes (10 hommes et 10 femmes). Quelques exemples de photos sont donnés en figure 4.3, et nous nommons cette base HFS_{CS} (pour HFS Compétence / Sympathie). Dans cette base, les visages présentent différentes expressions, coupes de cheveux, accessoires ou vêtements. Les arrière-plans sont également très variés et la qualité globale des images est également variable (visages flous, trop ou trop peu éclairés, etc.). Nous avons choisi de créer une base dont les photos sont les plus variées possibles, ce qui permet de tester l'influence de différents paramètres.

Dans deux expériences distinctes, nous avons demandé à 25 participants de juger le niveau de compétence ou de sympathie apparent de la personne sur les photos, et ce sur une échelle allant de 1 (incompétente/antipathique) à 6 (très compétente/sympathique). Les conditions de visionnage sont les mêmes pour tous les participants (distance à l'écran, luminosité de l'écran et de la pièce) et les images sont présentées dans un ordre aléatoire après une phase d'apprentissage durant laquelle les participants évaluent des images qui ne font pas partie de la base. La distribution des scores pour chaque évaluation est donnée sur les figures 4.4 et 4.5. Les moyennes et écarts-types de ces distributions sont respectivement de 3,31 et 0,47

FIGURE 4.3 – Exemples de photos de visage pour 4 personnes différentes de la base HFS_{CS}, pour lesquels nous avons collecté les scores moyens de compétence et de sympathie.

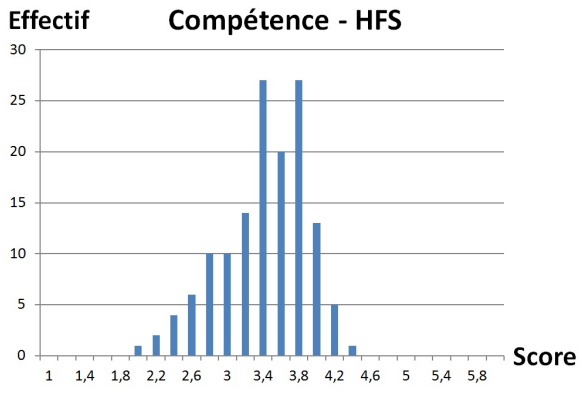


FIGURE 4.4 – Répartition des scores de compétence pour la base HFS_{CS}.

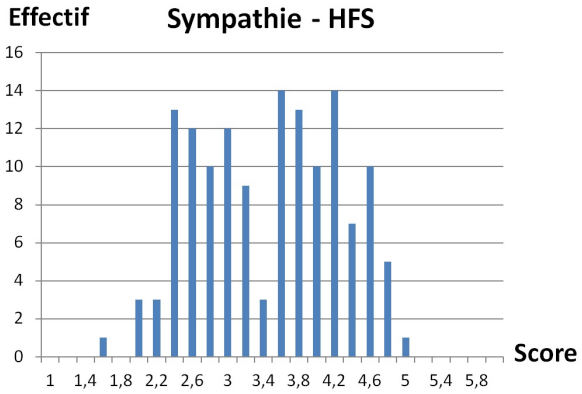


FIGURE 4.5 – Répartition des scores de sympathie pour la base HFS_{CS}.

pour la compétence, et de 3,37 et 0,79 pour la sympathie. Ces valeurs d'écart-type suggèrent que les participants ont eu plus de facilité à distinguer des visages très sympathiques ou très antipathiques : la présence de sourire sur les photos est très discriminante. A l'inverse, évaluer l'impression de compétence dégagée par la photo semble être un problème plus complexe car les participants ne s'éloignent que rarement du score moyen (écart-type des scores faible).

Sur les photos présentées en figure 4.3, un biais important peut être constaté. En effet, certains hommes portent un costume sur les photos. Nous n'avons pas considéré cet élément dans nos caractéristiques car nous n'avons pas d'outil permettant de décrire automatiquement le type de vêtement porté. Or il semble cohérent de penser que les vêtements influent fortement sur l'impression de compétence dégagée par une photo de visage. Les travaux de [Mazza et al. 2015] montrent en effet que le type de vêtement porté ainsi que l'endroit où est prise la photo (environnement intérieur/extérieur) sont des éléments importants pour qu'une photo soit utilisée dans un cadre professionnel. La prise en compte de ces informations est donc une piste d'amélioration des résultats présentés dans ce document.

4.3 Extraction des caractéristiques

Afin d'estimer automatiquement les impressions suggérées par une photo de visage, nous utilisons 3 catégories de caractéristiques. Dans la suite de ce chapitre, nous désignons chaque catégorie par un numéro de lot : Lot 1, Lot 2 et Lot 3. Les deux premiers lots correspondent à des caractéristiques utilisées couramment dans l'état de l'art, tandis que la troisième correspond à nos attributs de haut niveau.

4.3.1 Filtres de Gabor - Lot 1

De nombreuses méthodes, en particulier celle décrite dans les travaux de [Lajevardi et Lech 2008], exploitent les filtres de Gabor pour obtenir une estimation de l'expression faciale prédominante dans une image. Ceci se justifie par le fait que ces filtres permettent de caractériser les contours d'une image selon des orientations et des fréquences particulières. Ces contours et leurs orientations décrivent alors les déformations du visage (forme des yeux, de la bouche) responsables des expressions faciales. Si l'objectif de ce travail n'est pas l'évaluation d'expressions faciales, ces dernières sont largement corrélées aux impressions dégagées par une photo de visage. Pour résumer, nous justifions l'utilisation de ces filtres par le fait qu'un visage véhiculant une impression de sympathie (par exemple) n'aura pas les mêmes déformations des contours de la bouche ou des yeux qu'un visage dégageant une impression de compétence.

4.3.1.1 Définition des filtres de Gabor

Les filtres de Gabor sont des filtres linéaires utilisés pour la détection de contours. La réponse impulsionnelle de ces filtres est composée d'une sinusoïde complexe modulée par une enveloppe gaussienne. Pour le pixel (x,y) d'une image, l'expression de ces filtres peut être formulée par :

$$g\left(x,y;\lambda,\theta,\psi,\sigma,\gamma\right) = \exp\left(-\frac{x'^2+\gamma y'^2}{2\sigma^2}\right)\exp\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \tag{4.1}$$

où :

— $x' = x\cos(\theta) + y\sin(\theta)$,
— $y' = -x\sin(\theta) + y\cos(\theta)$,
— λ est la longueur d'onde du terme sinusoïdal, représentant la résolution,
— θ est l'orientation des bandes parallèles produites par le terme gaussien,
— ψ est le déphasage entre la partie réelle et imaginaire du terme sinusoïdal,
— σ est l'écart-type du terme gaussien,
— γ est le facteur de symétrie entre les axes x et y.

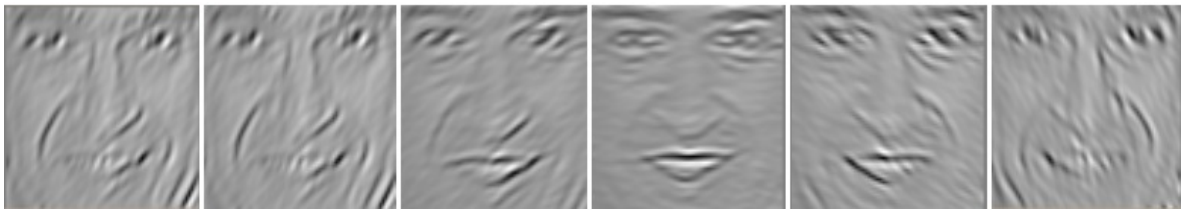


FIGURE 4.6 – Exemple des résultats obtenus par filtrage pour les 6 orientations considérées.

4.3.1.2 Extraction des filtres de Gabor

Dans nos travaux, l'extraction des informations obtenues par les filtres de Gabor ne se fait que sur les régions du visage. Pour cela, nous procédons d'abord à la détection du visage à l'aide de l'algorithme de Viola-Jones. En outre, les images sont prétraitées par une égalisation d'histogramme afin de s'assurer qu'elles exploitent la dynamique complète du canal des niveaux de gris.

Les paramètres suivants sont utilisés et recommandés par [Lajevardi et Lech 2008]. σ est fixé proportionnellement aux dimensions de l'image ; il est recommandé de choisir une valeur égale à 2π pour une image de dimensions 256×256. ψ est choisi nul, pour ne pas entraîner de déphasage. γ est pris égal à 1 afin de ne pas favoriser un axe particulier de l'image dans la mise en évidence des contours. Pour évaluer différentes orientations à différentes échelles, 6 orientations couvrant l'intégralité du cercle trigonométrique sont considérées : $0,\pi/6,2\pi/6,3\pi/6,4\pi/6,5\pi/6$. Les valeurs de λ utilisées dans ce travail sont 2, 4 et 8. Au total, l'image est ainsi filtrée 18 fois, pour 6 orientations et 3 échelles différentes. La figure 4.6 présente les résultats de différents filtrages pour $\lambda = 2$, et pour les 6 valeurs possibles de l'orientation du filtre.

Nous procédons ensuite de la même manière que Lajevardi et Lech : les 18 images obtenues par filtrages sont moyennées de manière à obtenir une seule carte contenant tous les contours et toutes les orientations. Enfin, la carte obtenue est sous-échantillonnée régulièrement de manière à obtenir un jeu de 192 caractéristiques décrivant les traits du visage. Chaque caractéristique correspond en effet à la valeur d'un pixel sur l'image des filtres fusionnés : une valeur élevée correspond à la présence de contours au niveau du pixel.

4.3.2 Position de points de repère sur le visage - Lot 2

Tout comme les descripteurs définis en 4.3.1.2 obtenus par la fusion de filtres de Gabor, les positions de points de repère sur le visage renseignent sur les déformations du visage responsables des expressions faciales. Les distances et angles formés par plusieurs points adjacents informent également sur ces déformations.

Les positions de certains points de repère du visage sont extraites à l'aide de deux outils disponibles sous la forme de services web : SkyBiometry et Betaface. Ces deux outils extraient automatiquement diverses informations sur les photos :

— Détection du ou des visages dans la photo,

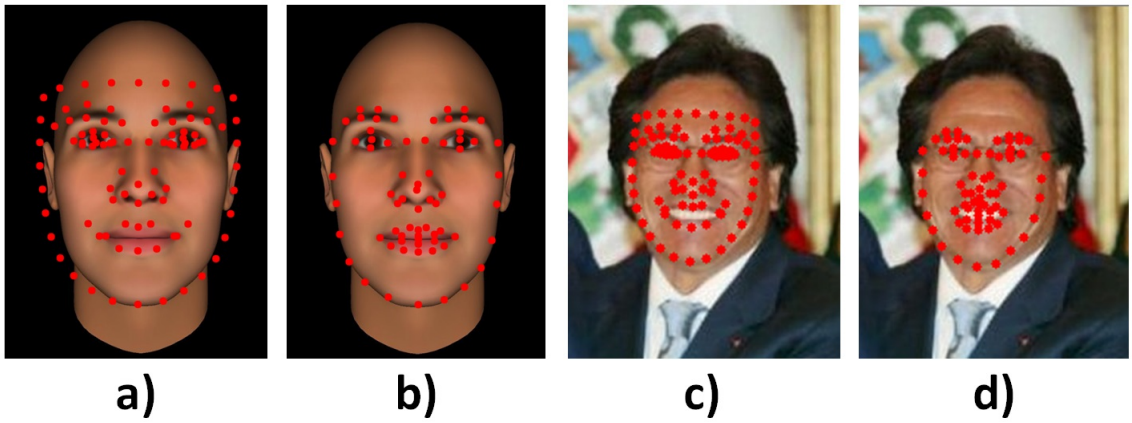


FIGURE 4.7 – Exemples des positions des points de repère pour des images synthétiques et réelles. a) et c) présentent les points fournis par Betaface, b) et d) ceux fournis par SkyBiometry.

— Localisation des points de repères de chaque visage,
— Calcul d'informations de haut niveau telles que la présence de sourire (voir 4.3.3).

Afin d'utiliser ces outils, il est nécessaire d'envoyer les photos à un serveur distant, qui renvoie à son tour la liste des visages détectés, leur position, ainsi qu'une liste de tous les éléments calculés dans chaque visage. Nous nous intéressons dans un premier temps aux positions des points de repère du visage. Chacun des deux services renvoie une liste de points de repère, où chaque point est défini par un identifiant, une abscisse et une ordonnée. Des exemples de position des points de repère pour différentes images (synthétiques et réelles) sont donnés sur la figure 4.7.

Au total, SkyBiometry propose l'extraction de 73 points de repère, tandis que Betaface extrait 94 points. Cette différence s'explique par le fait que Betaface définit des points de manière plus dense (par exemple 9 points dans chaque œil, contre seulement 5 pour SkyBiometry) et calcule également les contours supérieurs du visage, définis par la ligne de points au-dessus des sourcils sur la figure 4.7. Comme cela peut être observé sur cette figure, les deux outils ne sont pas infaillibles et la position de certains points n'est pas toujours exacte (voir les contours du visage **a** de la figure 4.7). La plupart des points de repère sont tout de même correctement localisés et nous n'étudions pas les performances de détection de ces outils dans notre travail.

Dans toutes les expériences décrites dans ce chapitre, nous utilisons les points de repère renvoyés par ces outils, et nous considérons que les positions sont correctes. Lorsqu'un point de repère est mal localisé, du bruit est introduit dans les données. Nous ne gérons pas les erreurs de position introduites par les outils de détection automatique.

4.3.3 Attributs de haut niveau - Lot 3

De la même manière que nous récupérons les positions des points de repère du visage grâce à Betaface et SkyBiometry, nous utilisons ces deux outils afin d'obtenir des informations de haut niveau sur le visage détecté. Nous utilisons également un troisième outil, SHORE, qui ne fournit pas les positions des points de repère. Il est tout à fait possible que ces outils fassent des erreurs.

Chacun des 3 outils permet de mesurer un certain nombre de valeurs. Ces valeurs peuvent être discrètes (est-ce un homme ou une femme ?) ou continues (quel âge a cette personne ?). Il est également possible qu'une même caractéristique présente deux composantes, l'une discrète et l'autre continue : cette personne sourit-elle ? Quelle est l'intensité du sourire ? Dans ce dernier cas, deux valeurs sont intégrées dans le vecteur décrivant l'image. La première définit ainsi la présence ou l'absence de sourire (0 ou 1) tandis que la seconde traduit l'intensité du sourire (sur une échelle de 0 à 1), ou dans certains cas la probabilité que le visage soit souriant. Ce mélange de caractéristiques continues et discrètes dans nos modèles est possible car nous avons construit notre cadre de travail (voir chapitre 2) dans ce but. En effet, notre implémentation de l'algorithme Relief ainsi que les différents algorithmes d'apprentissage sont tout à fait capables de gérer conjointement les deux types d'informations.

Au total, nous extrayons 63 valeurs différentes sur chaque image. Parmi ces 63 valeurs, 37 sont fournies par Betaface, 20 par SkyBiometry et 6 par SHORE. Une liste détaillée de ces attributs est donnée dans le tableau 4.1. Différents attributs, tels que le sexe ou l'âge d'une personne, informent sur la personne représentée sur la photo. Des informations sur les expressions faciales (Sourire, Émotions), ainsi que la présence d'attributs non permanents (Barbe, Moustache, Lunettes) sont également fournies. Les catégories définies par "Yeux, Bouche" traduisent des informations sur l'état de ces éléments (ouvertures de la bouche, des yeux), et celles définies par "Sourcils, Nez" les formes et positions de ces éléments (la position des sourcils renseigne par exemple sur les émotions de la personne). D'autres attributs, tels la coupe de cheveux ou la forme globale du visage sont également pris en compte. Le tableau 4.2 résume les différentes catégories d'attributs étudiés.

Tableau 4.1 – Liste des attributs utilisés pour l'évaluation des impressions de compétence et de sympathie.

Nom	Type	Valeurs possibles
Betaface		
Sexe	Discrète	$[\![0,1]\!]$: Femme ou homme
Sexe	Continue	[0,1] (confiance)
Âge	Continue	[0,1] : Entre 0 et 100 ans
Éthnicité	Discrète	$[\![0,6]\!]$: Asiatique (M-orient), Asiatique, Afro-Américain, Hispanique, Blanc, Moyen-oriental, Autre
Éthnicité	Continue	[0,1] (confiance)

Nom	Type	Valeurs possibles
Barbe	Discrète	⟦0,1⟧ : Absence, Présence
Barbe	Continue	[0,1] (confiance)
Barbe (taille)	Discrète	⟦0,2⟧ : Absence, Fine, Épaisse
Moustache	Discrète	⟦0,1⟧ : Absence, Présence
Moustache	Continue	[0,1] (confiance)
Moustache	Discrète	⟦0,2⟧ : Absence, Fine, Épaisse
Sourire	Discrète	⟦0,1⟧ : Absence, Présence
Sourire	Continue	[0,1] (confiance)
Sourire	Discrète	⟦0,1⟧ : Dents non visibles, Dents visibles
Lunettes	Discrète	⟦0,1⟧ : Absence, Présence
Lunettes	Continue	[0,1] (confiance)
Lunettes	Discrète	⟦0,1⟧ : Absence, Présence de monture
Sourcils (coins)	Discrète	⟦0,4⟧ : Très bas, Bas, Milieu, Relevés, Très relevés
Sourcils (posit.)	Discrète	⟦0,4⟧ : Très bas, Bas, Milieu, Hauts, Très hauts
Sourcils (taille)	Discrète	⟦0,4⟧ : Très gros, gros, Moyen, fins, Très fins
Yeux (coins)	Discrète	⟦0,4⟧ : Très bas, Bas, Milieu, Relevés, Très relevés
Yeux (distance)	Discrète	⟦0,4⟧ : T. éloignés, Éloignés, Moyen, Proches, T. proches
Yeux (position)	Discrète	⟦0,4⟧ : Très bas, Bas, Milieu, Hauts, Très hauts
Yeux (forme)	Discrète	⟦0,4⟧ : Très ronds, Ronds, Moyens, Fins, Très fins
Nez (forme)	Discrète	⟦0,4⟧ : Très droit, Droit, Moyen, Triang., Très triang.
Nez (largeur)	Discrète	⟦0,4⟧ : Très large, Large, Moyen, Fin, Très fin
Bouche (coins)	Discrète	⟦0,4⟧ : Très bas, Bas, Milieu, Relevés, Très relevés
Bouche (hauteur)	Discrète	⟦0,4⟧ : Très épaisse, Épaisse, Moyenne, Fine, Très fine
Bouche (largeur)	Discrète	⟦0,4⟧ : Très large, Large, Moyens, Petite, Très petite
Cheveux (coul.)	Discrète	⟦0,6⟧ : Noirs, Blonds, Roux, Bruns, Châtains, Lumière ou Couleur non naturelle
Cheveux (frange)	Discrète	⟦0,1⟧ : Absence, Présence
Cheveux (long.)	Discrète	⟦0,5⟧ : Aucun, T. courts, Courts, Moyens, Longs, T. longs
Cheveux (côtés)	Discrète	⟦0,3⟧ : Très fins, Fins, Moyens, Épais
Cheveux (haut)	Discrète	⟦0,4⟧ : Très courts, Courts, Moyens, Épais, Très épais
Visage (forme)	Discrète	⟦0,4⟧ : Très ronde, Ronde, Moyenne, Carrée, Très carrée
Visage (largeur)	Discrète	⟦0,1⟧ : Très fine, Fine, Moyenne, Large, Très large
Visage (joues)	Discrète	⟦0,4⟧ : T. larges, Larges, Moyennes, Petites, T. petites
SkyBiometry		
Sexe	Discrète	⟦0,1⟧ : Femme ou homme
Sexe	Continue	[0,1] (confiance)
Yeux	Discrète	⟦0,1⟧ : Ouverts, Fermés
Yeux	Continue	[0,1] (confiance)
Lèvres	Discrète	⟦0,1⟧ : Ouvertes, Fermées
Lèvres	Continue	[0,1] (confiance)
Sourire	Discrète	⟦0,1⟧ : Absence, Présence

Nom	Type	Valeurs possibles
Sourire	Continue	[0,1] (confiance)
Lunettes	Discrète	⟦0,1⟧ : Absence, Présence
Lunettes	Continue	[0,1] (confiance)
Lunettes soleil	Discrète	⟦0,1⟧ : Absence, Présence
Lunettes soleil	Continue	[0,1] (confiance)
Émotion	Discrète	⟦0,6⟧ : Neutre/Joie/Tristesse/Colère/Surprise/Dégoût/Peur
Neutre	Continue	[0,1] : Niveau d'expression
Joie	Continue	[0,1] : Niveau d'expression
Tristesse	Continue	[0,1] : Niveau d'expression
Colère	Continue	[0,1] : Niveau d'expression
Surprise	Continue	[0,1] : Niveau d'expression
Dégoût	Continue	[0,1] : Niveau d'expression
Peur	Continue	[0,1] : Niveau d'expression
SHORE		
Sexe	Discrète	⟦0,1⟧ : Femme ou homme
Âge	Continue	[0,1] (entre 0 et 100 ans)
Œil gauche	Discrète	⟦0,1⟧ : Ouvert, Fermé
Œil droit	Discrète	⟦0,1⟧ : Ouvert, Fermé
Bouche	Discrète	⟦0,1⟧ : Ouverte, Fermée
Joie	Continue	[0,1] : Niveau d'expression

Tableau 4.2 – Liste simplifiée des catégories d'attributs calculés par chaque outil. Le nombre de valeurs (discrètes et continues) décrivant chaque catégorie est indiqué dans chaque case du tableau.

	Sexe	Âge	Éthnicité	Sourire	Émotions	Barbe	Moustache	Lunettes	Yeux	Bouche	Sourcils	Nez	Cheveux	Visage
Betaface	2	1	2	3		3	3	3	4	3	3	2	5	3
SkyBiometry	2			2	8			4	2	2				
SHORE	1	1			1				2	1				

Nous observons dans le tableau 4.2 que les différents outils ne calculent pas les mêmes attributs. La plupart évaluent l'âge, le sexe ou les expressions faciales dans les visages, ainsi que des éléments tels que l'ouverture de la bouche ou des yeux (catégories "Bouche, Yeux"). Les 3 outils présentés, très récents dans les cas de Betaface et SkyBiometry (les deux outils existent respectivement depuis 2013 et 2012), ne détaillent pas la façon dont les informations sont calculées. D'autres outils en ligne, tels que FacePlusPlus (plus récent encore), extraient le même type d'informations et nous avons choisi de nous limiter à 3 outils car les algorithmes d'estimation des attributs reposent généralement sur les mêmes méthodes.

Remarquons que les attributs présentés ne sont pas suffisants pour fournir un modèle exhaustif des impressions véhiculées par des photos de visage, et certains éléments sont manquants. Par exemple, il n'y a dans ces attributs aucune information sur l'arrière-plan de l'image, contenant des informations discriminantes : un arrière-plan simple dégage un air plus

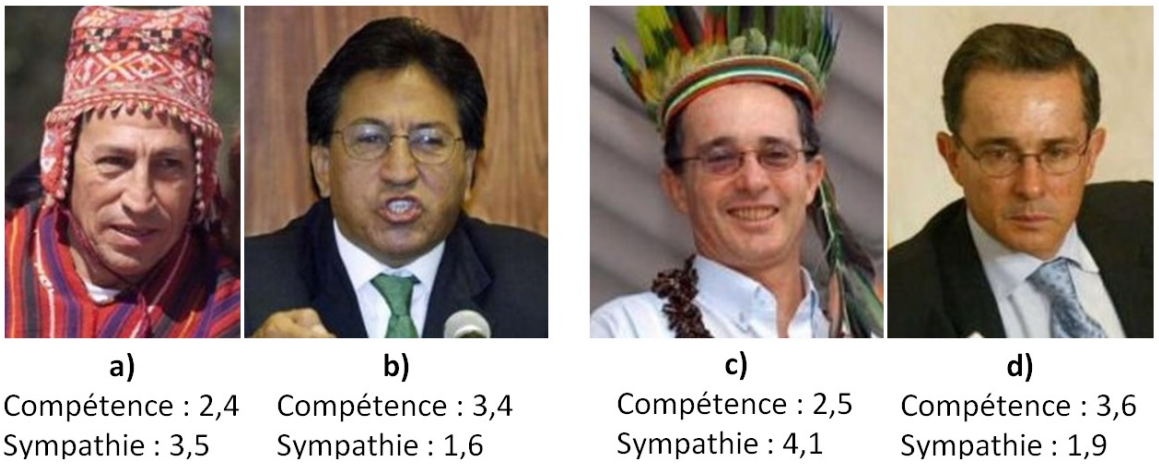


FIGURE 4.8 – Exemples de 4 photos de visage pour deux personnes différentes. Les photos a) et c) ont des scores de vérité terrain élevés pour la sympathie, les photos b) et d) ont des scores de compétence élevés. L'échelle des scores va de 1 à 6 (voir expérience décrite en 4.2).

sérieux qu'un fond trop chargé ou contenant des éléments inattendus/inappropriés. De plus, aucune information sur les vêtements n'est prise en compte : un homme en costume véhicule automatiquement une impression de compétence, tandis que la présence d'un couvre-chef amusant ou d'une tenue exotique a tendance à véhiculer de la sympathie, comme le montrent les photos présentées en figure 4.8. D'autres facteurs modifient les impressions dégagées par ces photos, en particulier les expressions faciales sont importantes pour l'évaluation de la sympathie. La présence, l'absence ou la quantité de maquillage, de bijoux ou de pilosité faciale sont également des éléments à prendre en compte.

4.4 Étude de l'influence des caractéristiques

Dans cette partie, nous testons séparément chaque lot de caractéristiques pour la classification des images synthétiques en 7 niveaux d'expression de compétence et de sympathie. Ces bases présentent en effet des traits caricaturaux et nous espérons obtenir des informations sur les caractéristiques les plus discriminantes. Dans cette section, nous montrons que même dans le cas d'images caricaturales (pas de cheveux, de lunettes, de barbe) dont les traits sont contrôlés et exagérés, les attributs que nous extrayons à l'aide des trois outils (Betaface, SkyBiometry, SHORE) permettent d'obtenir des performances similaires aux filtres de Gabor et aux points de repère du visage.

Nous étudions ensuite à l'aide de l'algorithme Relief quels sont les points de repère dont la position influe sur les impressions dégagées, et quels sont les attributs de haut niveau les plus discriminants.

FIGURE 4.9 – Performances de classification sur les visages synthétiques pour l'estimation de la compétence pour chaque lot de caractéristique. K représente l'amplitude des erreurs de classification.

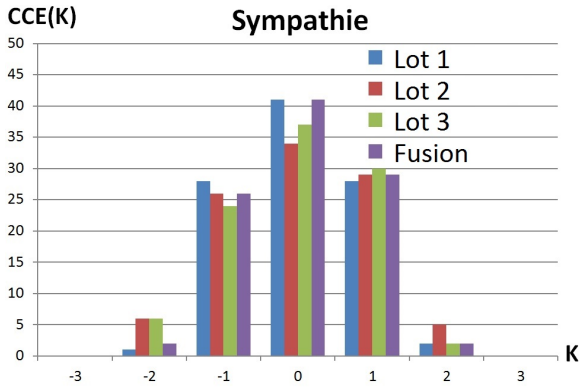


FIGURE 4.10 – Performances de classification sur les visages synthétiques pour l'estimation de la sympathie pour chaque lot de caractéristique. K représente l'amplitude des erreurs de classification.

4.4.1 Comparaison des 3 lots de caractéristiques

Les filtres de Gabor (Lot 1) sont largement utilisés pour la reconnaissance d'expressions faciales [Lajevardi et Lech 2008], elles-mêmes significativement corrélées à l'évaluation des traits de caractère : un visage joyeux dégage par exemple une impression de sympathie. Les précédents travaux sur l'évaluation des traits de caractère reposent également sur l'analyse des positions des points de repère [Rojas et al. 2010 ; Rojas et al. 2011] (Lot 2). Nous testons ainsi la pertinence des points de repère fournis par chacun des outils (Betaface et SkyBiometry) pour la classification des visages synthétiques. Enfin, nous étudions les résultats que nous obtenons à l'aide des attributs de haut niveau (Lot 3). L'utilisation de ces attributs dans le cadre de l'évaluation de traits de caractères induits par une photo de visage est l'apport principal de nos travaux par rapport à l'état de l'art.

Nous considérons dans ces premières expériences uniquement l'algorithme GBT, qui s'est révélé être le plus stable de nos algorithmes lorsque les données sont de nature très différentes (grande dimension, valeurs discrètes et continues). A l'aide de cet algorithme, nous obtenons les résultats présentés sur les figures 4.9 et 4.10, représentant les performances de classification obtenues respectivement pour l'estimation de la compétence et de la sympathie.

Ces figures montrent que des résultats très proches sont obtenus pour les différentes méthodes, les caractéristiques de Gabor étant légèrement plus discriminantes. En outre, nous observons que les performances n'augmentent que très légèrement lorsque nous fusionnons les 3 lots de caractéristiques. En effet, nous montrons en section 4.4.2 que les informations obtenues par les filtres de Gabor ou par les positions des points de repère (lots 1 et 2) sont semblables aux informations fournies par les attributs de haut niveau (lot 3).

4.4.2 Caractéristiques discriminantes

L'utilisation de l'algorithme Relief nous permet d'étudier quelles sont les caractéristiques les plus discriminantes. Pour des raisons de représentation, nous choisissons de nous limiter ici à l'étude des lots 2 (les points de repère peuvent être affichés sur les photos de visage) et 3 (l'interprétation des attributs est immédiate).

Les points de repère détectés par SkyBiometry dont les valeurs obtenues par l'algorithme Relief sont les plus élevées sont affichés sur la figure 4.11. Nous voyons que les contours des sourcils et de la bouche sont importants pour les deux traits de caractère. Ceci s'explique par le fait que ces points sont directement liés aux expressions faciales, et un parallèle entre ces points et les attributs de haut niveau les plus significatifs (voir tableau 4.3) peut être fait : la position des sourcils ou la présence d'un sourire sont de bons indicateurs des impressions de compétence ou de sympathie véhiculées par les images. Dans le cas de la sympathie, les points définissant les contours du visage sont également pris en compte : il est possible que leur position informe sur la forme du visage : un visage rond est généralement perçu comme étant plus sympathique. Enfin, les points de repère décrivant les yeux semblent largement pris en compte dans le cas de l'évaluation de la compétence ; encore une fois cette information est également présente dans les attributs de haut niveau discriminants du tableau 4.3 (Forme des yeux).

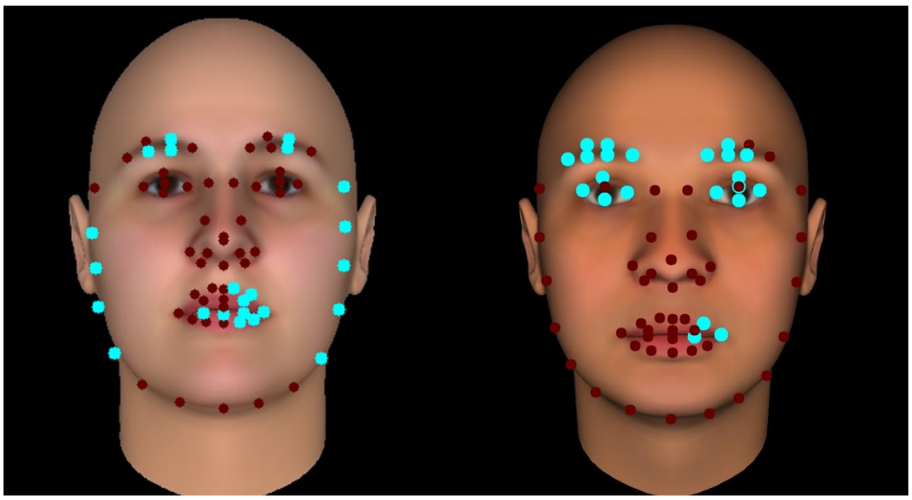


FIGURE 4.11 – Les points de repère pertinents selon l'algorithme Relief sont visibles en bleu clair. A gauche sont présentés les points caractéristiques de la sympathie, à droite ceux caractéristiques de la compétence.

Les avantages liés à l'utilisation de nos attributs sont donc multiples. Ils permettent en effet d'ajouter de l'information aux modèles, tout en fournissant des informations très faciles à interpréter pour des humains.

TABLEAU 4.3 – Liste des attributs les plus discriminants selon l'algorithme Relief, pour l'évaluation de la compétence et de la sympathie. Les caractéristiques sont listées par ordre d'importance, et les lettres B (pour Betaface), S_k (pour SkyBiometry) et S_H (pour SHORE) désignent les différents outils permettant de calculer les attributs.

Critère	Caractéristiques discriminantes
Compétence	Émotion (S_k), Joie (S_k, S_h), Forme des yeux (B), Forme de la bouche (B), Expression neutre (S_k), Taille des sourcils (B), Position des sourcils (B), Position des yeux (B).
Sympathie	Sourire (S_k, S_h), Hauteur de la bouche (B), Joie (S_k, S_h), Position des sourcils (B), Colère (S_k), Largeur de la bouche (B), Position des yeux (B), Coins de la bouche (B).

4.4.3 Comparaison des outils d'extraction des attributs

Nous comparons maintenant les performances de nos 3 outils d'extraction d'attributs. En effet, chaque outil extrait des caractéristiques différentes : Betaface fournit de nombreuses informations sur la forme du visage, des yeux ou de la bouche, tandis que SkyBiometry nous renseigne précisément sur les expressions faciales. Enfin, SHORE fournit des informations plus synthétiques : l'âge, le sexe, l'ouverture de la bouche ou des yeux, l'expression de la joie. Il semble donc pertinent de comparer les performances de chaque outil afin de voir si les informations fournies par SHORE sont suffisantes, ou s'il est nécessaire d'intégrer des informations relatives aux expressions faciales (SkyBiometry) ou à la forme du visage (Betaface).

Les bases d'images synthétiques sont ainsi testées en utilisant à chaque fois les caractéristiques fournies par un seul outil (soit une seule ligne du tableau 4.2). Les résultats de classification sont présentés pour chaque ensemble de caractéristiques dans le tableau 4.4. Pour chaque outil, nous observons que les résultats sont significativement supérieurs au hasard. En outre, les résultats obtenus par Betaface et SkyBiometry sont à peu près équivalents, tandis que les attributs proposés par SHORE sont moins discriminants. Ceci s'explique par l'absence d'un grand nombre d'attributs pour ce dernier outil (seulement 6 valeurs décrivant l'image) : pas d'information sur la présence de sourire ou la position des sourcils par exemple. Il apparaît également que les attributs proposés par Betaface sont plus efficaces pour évaluer la compétence, tandis que ceux proposés par SkyBiometry évaluent plus précisément la sympathie. Ceci peut en partie s'expliquer par la présence d'attributs décrivant les émotions pour SkyBiometry (un visage joyeux est souvent associé à un visage sympathique), et d'attributs tels que l'âge et surtout la forme du visage pour Betaface (forme des yeux, des sourcils, du visage).

TABLEAU 4.4 – Comparaison des performances de classification des attributs fournis par chacun des 3 outils, évaluées par le taux de bonne classification T_{BC} et l'erreur MCE en %, définie dans l'équation 2.16.

Outil	Compétence		Sympathie	
	T_{BC}	MCE	T_{BC}	MCE
Betaface	**36**	**34**	30	40
SkyBiometry	35	**34**	**36**	**34**
SHORE	25	53	19	58

4.5 Estimation des impressions de compétence et de sympathie pour une personne donnée

Notre objectif est de sélectionner automatiquement les photos véhiculant les impressions de sympathie ou de compétence les plus élevées pour une personne donnée. Nous cherchons ainsi à prédire un score à chaque photo correspondant à une personne ; il sera alors possible de conserver uniquement celles dont le score de prédiction est le plus élevé. Ces expériences sont faites sur la base HFS_{CS}.

Rappelons que la base HFS_{CS} contient 140 photos de 20 personnes différentes. Le protocole de validation croisée utilisé jusqu'ici implique l'utilisation de photos de tests aléatoires ; il y a donc de très fortes chances pour que des photos différentes d'une personne donnée se retrouvent à la fois dans les bases d'apprentissage et de test. Ceci introduit un biais lors de la prédiction : les algorithmes vont évaluer les photos après avoir appris un modèle à partir de photos de la personne à évaluer. Pour éviter ce phénomène, nous effectuons la validation croisée en apprenant un modèle sur 19 personnes, puis en testant sur la personne restante. Cela revient à faire une validation croisée à 20 groupes.

Pour chaque expérience, nous sélectionnons les 40 meilleures caractéristiques selon l'algorithme Relief, l'apprentissage est ensuite effectué par l'algorithme de fusion LSF. En procédant à une analyse similaire à celle proposée au chapitre 3, section 3.5, nous avons en effet observé que le fait de conserver 40 des 63 attributs de haut niveau permet généralement d'obtenir les résultats les plus précis.

4.5.1 Classification binaire

Les performances de classification obtenues pour la base HFS_{CS} à l'aide du protocole que nous venons de décrire sont données dans le tableau 4.5. Nous n'utilisons ici que les attributs de haut niveau (Lot 3) car les autres caractéristiques se sont révélées très peu efficaces dans ces expériences, légèrement au-dessus du hasard pour la sympathie ($57{,}6 \pm 4{,}1\%$) et au niveau du hasard pour la compétence.

Nous observons dans le tableau 4.5 que les résultats sont très différents selon l'objectif :

TABLEAU 4.5 – Performances de classification sur la base HFS_{CS}, pour l'évaluation de la compétence et de la sympathie d'une personne donnée. Les performances sont mesurées par le taux de bonne classification (T_{BC}) en %. Seuls les attributs de haut niveau (Lot 3) sont considérés.

Impression dégagée	Compétence	Sympathie
T_{BC} (%)	$74{,}9 \pm 2{,}4$	$61{,}5 \pm 2{,}9$

TABLEAU 4.6 – Performances de régression sur la base HFS_{CS}, pour l'évaluation de la compétence et de la sympathie d'une personne donnée. Les performances sont mesurées par les corrélation de Pearson (R) et de Spearman (ρ) en %. "Lot 1" désigne l'utilisation des filtres de Gabor uniquement, "Lot 2" l'utilisation des attributs de points de repère, "Lot 3" l'utilisation des attributs de haut niveau et "Fusion" l'utilisation conjointe des deux types de caractéristiques.

Caractéristiques	Lot 1	Lot 2	Lot 3	Fusion
Compétence (R)	$11{,}8 \pm 7{,}0$	$19{,}8 \pm 1{,}8$	$27{,}6 \pm 6{,}0$	**33,1±3,3**
Compétence (ρ)	$15{,}5 \pm 6{,}4$	$18{,}1 \pm 1{,}6$	$28{,}3 \pm 5{,}8$	**34,5±3,4**
Sympathie (R)	$36{,}7 \pm 0{,}3$	$47{,}4 \pm 4{,}3$	**72,1±1,0**	**74,7±1,0**
Sympathie (ρ)	$38{,}4 \pm 0{,}3$	$47{,}9 \pm 4{,}4$	**72,6±1,0**	**74,4±0,9**

l'algorithme présente plus de facilités à classer les photos par sympathie que par compétence. Les performances de classification concernant la compétence sont par ailleurs tout de même significativement au-delà du hasard lorsque les attributs de haut niveau sont considérés. Cette différence peut s'expliquer par le fait que dans le cas de la compétence, les scores de vérité terrain sont très proches du score médian (voir figure 4.4). Afin d'affiner les évaluations, il est possible d'effectuer de la régression.

4.5.2 Régression

En suivant le même protocole que pour la classification, nous obtenons les performances de régression indiquées dans le tableau 4.6. Cette fois, nous considérons également les lots de caractéristiques 1 et 2 afin de comparer les performances entre les différents lots.

Nous observons que la précision des prédictions d'impression de compétence est particulièrement faible. Une explication possible est le manque de certaines informations pertinentes encodées dans les attributs de haut niveau. Afin d'améliorer les performances d'évaluation de la compétence pour une personne donnée, il serait intéressant de prendre en compte des informations sur les vêtements et l'arrière-plan. En effet un grand nombre de photos présentes dans la base présentent des hommes habillés en costume, évalués comme étant compétents par les humains, alors que nous ne prenons pas cette information en compte actuellement. À l'inverse, Les évaluations de sympathie reposent surtout sur des critères liés aux émotions et expressions faciales, très largement encodées dans nos attributs de haut niveau : les performances de régression sont donc particulièrement élevées.

TABLEAU 4.7 – Performances de classification (2 catégories) obtenues sur 200 images synthétiques, pour l'évaluation de la compétence et de la sympathie. Les performances sont mesurées par le taux de bonne classification en %. "Lot 2" désigne l'utilisation des points de repère uniquement, "Lot 3" l'utilisation des attributs de haut niveau et "Fusion" l'utilisation conjointe des deux types de caractéristiques.

Méthode	Compétence			Sympathie		
	Lot 2	Lot 3	Fusion	Lot 2	Lot 3	Fusion
Rojas et al. + SVM	$64{,}3 \pm 3{,}9$	/	/	**66,7±6,1**	/	/
Ce travail + SVM	$64{,}4 \pm 1{,}4$	$63{,}8 \pm 1{,}4$	**70,0±2,0**	$65{,}9 \pm 2{,}4$	$57{,}3 \pm 3{,}5$	**66,7±2,9**
Ce travail + LSF	$64{,}3 \pm 2{,}2$	**69,2±2,1**	**70,6±1,3**	**67,8±1,7**	$60{,}0 \pm 2{,}4$	**68,0±2,7**

4.6 Comparaison avec l'état de l'art

A notre connaissance, il n'existe pas de travaux dans lesquels des expériences de régression ont été menées afin d'estimer les impressions de compétence ou de sympathie dégagées par une photo de visage. Nous nous limitons donc à des expériences de classification. Nous comparons dans un premier temps nos résultats à ceux de l'état de l'art sur les visages synthétiques, puis nous proposons une comparaison de nos travaux à l'état de l'art sur des photographies. Pour chaque expérience, seules les 40 caractéristiques les plus discriminantes selon l'algorithme Relief sont considérées.

4.6.1 Performances sur les visages synthétiques

Nous utilisons la base de 300 visages synthétiques évalués selon les impressions de compétence ou de sympathie décrite en 4.2.1. Afin de reproduire les expériences de classification décrites dans les travaux de [Rojas et al. 2010], nous procédons aux opérations suivantes :

- Seuls les tiers inférieurs et supérieurs des images sont conservés. Cela signifie que les 100 images à la compétence/sympathie la plus faible sont groupées dans une classe, les 100 aux scores les plus élevés dans une autre classe. L'objectif est de distinguer ces deux classes.
- Rojas et al. utilisent différents algorithmes d'apprentissage, dont SVM. Afin de comparer les performances des attributs uniquement, nous utilisons donc SVM dans nos expériences également. Nous utilisons ensuite l'algorithme LSF afin d'optimiser les performances.
- Rojas et al. utilisent des descripteurs de dimension 1134, caractérisant les positions de 21 points de repère, des distances ainsi que des angles entre ces points. Nous effectuons trois expériences : la première en utilisant les points de repère également (uniquement les indices de position), la seconde en utilisant les attributs de haut niveau, puis une troisième en fusionnant les données.

Les résultats sont rapportés dans le tableau 4.7.

Ce tableau montre que les attributs de haut niveau que nous calculons nous permettent

TABLEAU 4.8 – Performances de classification (2 catégories) obtenues sur 44 photos, pour l'évaluation de la confiance et de la dominance. Les performances sont mesurées par le taux de bonne classification en %. "Lot 2" désigne l'utilisation des points de repère uniquement, "Lot 3" l'utilisation des attributs de haut niveau et "Fusion" l'utilisation conjointe des deux types de caractéristiques.

Méthode	Confiance			Dominance		
	Lot 2	Lot 3	Fusion	Lot 2	Lot 3	Fusion
Rojas et al. + SVM	$65{,}2 \pm 16$	/	/	**77,7±25**	/	/
Ce travail + SVM	$61{,}3 \pm 3{,}0$	$63{,}4 \pm 2{,}1$	$62{,}9 \pm 2{,}9$	$70{,}0 \pm 3{,}5$	$67{,}7 \pm 3{,}1$	$70{,}9 \pm 2{,}7$
Ce travail + LSF	$61{,}7 \pm 3{,}4$	$63{,}6 \pm 3{,}5$	**72,1±3,2**	$71{,}8 \pm 3{,}7$	$68{,}1 \pm 3{,}5$	**78,0±4,3**

d'atteindre les mêmes performances que les points de repère sur les visages synthétiques, ce qui confirme les expériences proposées en 4.4.1. La fusion de ces deux types de caractéristiques permet d'ajouter de l'information pertinente et améliore les résultats, surtout pour l'évaluation de la compétence. Les performances d'évaluation de la sympathie sont ici limitées car les visages présentent tous des expressions quasiment neutres.

4.6.2 Performances sur les visages réels

Comme dans le cas des visages synthétiques, nous avons comparé nos jeux de caractéristiques avec ceux présentés dans les travaux de [Rojas et al. 2010]. La base utilisée pour les tests est Karolinska, pour laquelle les images sont évaluées selon les impressions de confiance et de dominance dégagées par les visages. Il est important de rappeler que la base ne contient que 66 images. L'objectif étant d'effectuer de la classification, seules les photos aux scores les plus extrêmes sont utilisées : le tiers inférieur correspond à la catégorie de photos de visage véhiculant un visage peu digne de confiance / peu dominant, le tiers supérieur aux photos dont le visage semble digne de confiance / dominant. Nous cherchons à reproduire les résultats proposés par Rojas et al. et obtenons le tableau 4.8.

Nos performances sont en moyenne similaires à celles obtenues par Rojas et al., et les résultats sont bien plus stables (voir les écart-types indiqués dans le tableau 4.8). Répéter l'expérience avec des ensembles d'apprentissage et de test différents modifie fortement les performances. Remarquons qu'une seule image mal classée induit une chute de performances de l'ordre de 2,5%. Il est donc difficile de tirer des conclusions de ces expériences, si ce n'est que l'utilisation d'attributs permet d'obtenir des résultats équivalents à ceux obtenus par les points de repère, alors que les photos sont conçues de manière à présenter des visages neutres, sans accessoire particulier : une part importante des attributs de haut niveau n'est pas utilisée (lunettes, moustache, barbe, expressions faciales). Les performances des algorithmes sont tout de même significativement au-dessus du hasard.

4.7 Fusion des modèles de qualité esthétique et de sympathie

Nous avons défini un modèle d'estimation de la qualité esthétique dans le chapitre 3, et nous venons de proposer un modèle d'évaluation de la sympathie dans ce chapitre. Ces deux modèles permettent d'obtenir des résultats de régression intéressants, avec des corrélations au-delà de 70% pour les deux évaluations sur la base HFS.

Nous rappelons qu'une des applications majeures de ce travail de thèse est la sélection automatique de photos de visage appropriées à un contexte d'utilisation particulier. Pour une personne donnée, nous cherchons à sélectionner automatiquement les photos dans lesquelles celle-ci présente un visage sympathique (souriant, joyeux) tout en proposant une photo de qualité esthétique élevée. Il deviendrait alors bien plus facile de trier automatiquement un dossier contenant un grand nombre de photos : suppression des ratés, sélection des meilleurs clichés.

4.7.1 Méthode de fusion

Afin de proposer une méthode d'évaluation de la qualité esthétique et de l'impression de sympathie, il est nécessaire d'évaluer conjointement les deux critères (qualité esthétique et sympathie). Cela signifie que dans le cadre d'utilisation d'algorithmes d'apprentissage supervisés, les bases de photos d'apprentissage doivent être évaluées selon les deux critères à la fois. Or il est difficile pour un humain d'évaluer directement une photo en tenant compte de deux critères très subjectifs simultanément (donc l'obtention d'une vérité terrain est délicate), et nous choisissons pour cela d'évaluer les deux critères séparément, puis de fusionner ces prédictions.

La base HFS_{CS} contient 140 photos évaluées selon les deux critères et dans les mêmes conditions, par deux groupes différents d'environ 25 personnes. Nous remarquons au passage qu'il existe une corrélation positive ($R = 0{,}22$) entre les scores de vérité terrain de qualité esthétique et d'impression de sympathie. Des expériences supplémentaires pourraient être menées afin de tester si l'impression de sympathie améliore la qualité esthétique, ou si une photo de qualité esthétique élevée augmente l'impression de sympathie. Dans le cas de photographies amateurs (comme celles incluses dans la base HFS_{CS}), il semble raisonnable de penser que la présence d'un sourire donne l'impression d'une photo réussie car une expression joviale est généralement attendue. A l'inverse, les photographies de professionnels tendent à présenter des visages plus neutres, il serait donc intéressant d'estimer l'impression de sympathie sur ce type de photographie afin de vérifier la présence ou non d'une corrélation entre qualité esthétique et impression de sympathie.

Afin d'obtenir un score global tenant compte de ces deux critères, plusieurs stratégies de fusion peuvent être proposées. Une première idée est de faire la moyenne des deux scores : une photo dont les scores de qualité esthétique et d'impression de sympathie sont moyens sera jugée comme étant moyenne dans le cadre de cette application. Toutefois, faire la moyenne des deux critères pose le problème suivant : si l'un des scores est très faible (par exemple une

qualité esthétique très médiocre) et l'autre très élevée (un très beau sourire), le score moyen sera tout de même moyen, or nous souhaitons qu'une valeur trop faible pour l'un des critères soit éliminatoire.

Pour cela, plutôt que de moyenner les scores, nous choisissons de les multiplier entre eux. Un score très faible pour l'un des critères induit alors automatiquement un score de fusion faible. Nous définissons la vérité terrain comme le produit des vérités terrain issues de chaque critère :

$$\text{Vérité terrain(Fusion)} = \text{Vérité terrain(Qualité esthétique)} \times \text{Vérité terrain(Sympathie)} \tag{4.2}$$

De même, la prédiction finale est définie comme la fusion des prédictions :

$$\text{Prédiction(Fusion)} = \text{Prédiction(Qualité esthétique)} \times \text{Prédiction(Sympathie)} \tag{4.3}$$

En employant cette méthode, nous obtenons les nuages de points présentés sur la figure 4.12. De la même manière que pour les estimations séparées de qualité esthétique ou de sympathie, il existe toujours des erreurs de prédiction importantes pour certaines images. Toutefois, nous remarquons que pour cet ensemble de photos (HFS_{CS}), il n'y a que très peu d'images ayant une valeur de vérité terrain très élevée (par exemple au-delà de 15) et une valeur de prédiction très faible (en-dessous de 10), et inversement. Cela signifie que les cas où une photo de très bonne qualité esthétique et véhiculant une impression de sympathie élevée est automatiquement supprimée sont très rares. De la même manière, une photo de très mauvaise qualité esthétique ou présentant un visage antipathique ne sera que très rarement sélectionnée par l'algorithme.

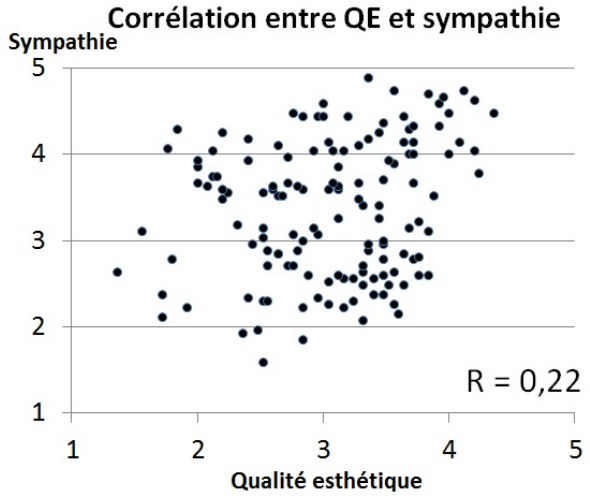
4.7.2 Exemples de sélection de photos pour une personne donnée

La fusion des critères d'évaluation citée dans la section précédente peut être appliquée à la sélection ou la suppression automatique de photos pour une personne donnée, comme cela est présenté sur la figure 4.13. Cette figure présente des seuils qu'il est possible de fixer manuellement en fonction du nombre de photos souhaitées : plus le seuil représenté en vert (Seuil Haut) est élevé, plus le nombre de photos automatiquement sélectionnées est faible. De même, plus le seuil rouge (Seuil Bas) est élevé, plus le nombre de photos automatiquement supprimées est élevé. Plutôt que de fixer un seuil, il est également possible de fixer un nombre de photos à conserver en demandant par exemple à l'algorithme de ne conserver que les 3 meilleures photos (meilleures au sens d'un score de prédiction élevé).

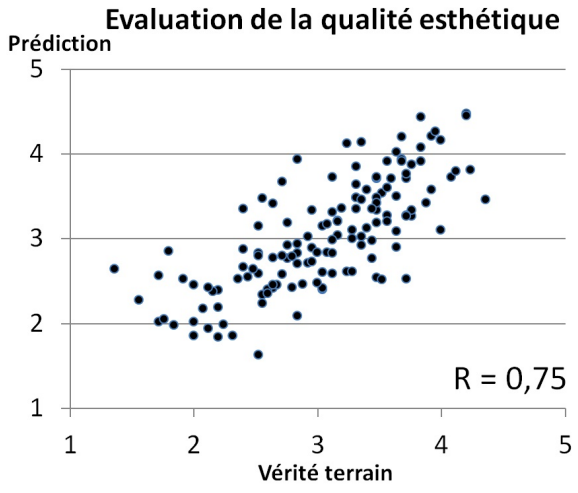
4.8 Conclusion

Limites du modèle

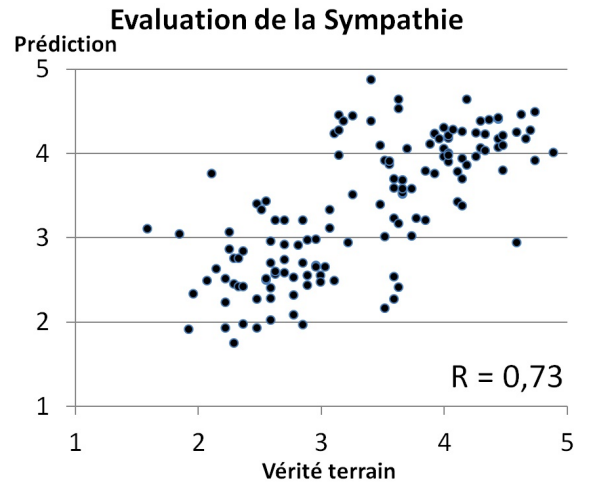
Nous avons vu que dans le cas de l'évaluation de la sympathie sur des images réelles, il est possible d'obtenir des performances de prédiction intéressantes grâce à l'utilisation de filtres



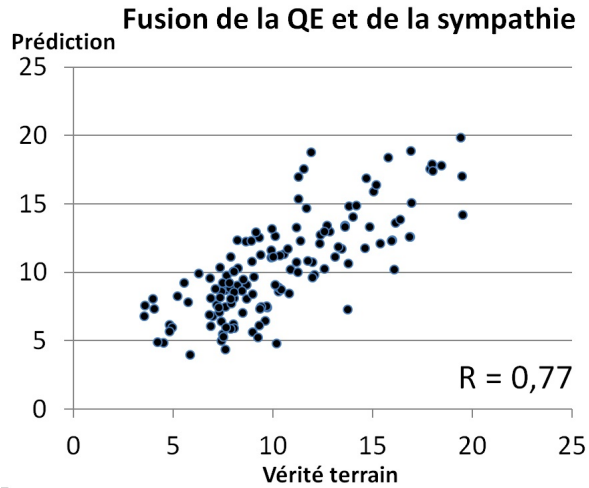
(a) Corrélation entre qualité esthétique et impression de sympathie.



(b) Nuage de points obtenu pour l'estimation de la qualité esthétique (QE).



(c) Nuage de points obtenu pour l'estimation de l'impression de sympathie.



(d) Nuage de points obtenu pour la fusion (produit) des vérités terrain et prédictions.

FIGURE 4.12 – Différents nuages de points obtenus pour la base HFS_{CS}.

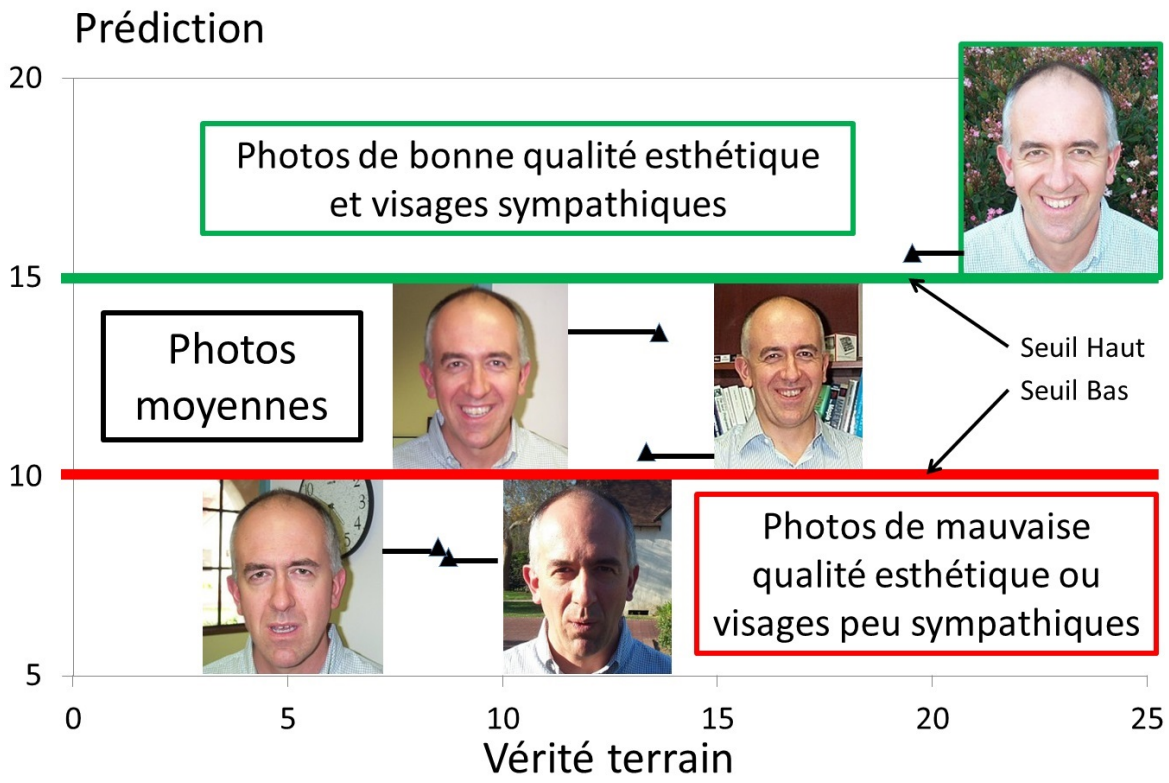


FIGURE 4.13 – Exemples de vérités terrain et de prédictions pour différentes photos d'une même personne. Les images au-dessus de la ligne verte sont automatiquement sélectionnées tandis que celles en-dessous de la ligne rouge sont supprimées.

de Gabor, mais significativement inférieures à celles obtenues par l'utilisation d'attributs de haut niveau. Toutefois, concernant l'évaluation de la compétence, les performances ne sont pas encore suffisantes pour exploiter les résultats en pratique. Au moins deux facteurs expliquent ce phénomène. Tout d'abord, ce critère est complexe à définir et hautement subjectif : la distribution des scores de compétence sur la base HFS_{CS} montre que les observateurs ne s'éloignent que très peu du score moyen, un consensus est difficile à obtenir. Si l'évaluation de la sympathie est largement corrélée à d'autres éléments tels que la présence d'un sourire, il n'existe pas d'attributs aussi discriminants pour la compétence. Le second problème rencontré pour l'évaluation de la compétence est le manque de certaines informations discriminantes, telles que le type de vêtements ou le type d'arrière-plan.

Une limite importante concernant ce travail est le manque de données annotées. La base de photos que nous avons constituée dans le cadre de ce travail ne regroupe que 140 photos, or il est difficile de créer des modèles généralisables à partir de si peu de photos. En effet, les critères considérés sont très variés, et toutes les catégories de personnes (hommes, femmes, différents âges) doivent être représentées. Une étape nécessaire à la création d'un modèle plus efficace est donc la création d'une base de photos plus conséquente.

Bilan

Nous avons proposé dans ce chapitre une méthode d'estimation automatique des impressions de compétence ou de sympathie véhiculées par des photos de visage. Pour cela, nous avons utilisé différentes informations sur l'image, encodées sous la forme d'attributs de haut niveau : la personne sourit-elle ? Est-ce un homme ou une femme ?

Nous avons montré que l'utilisation de ces attributs permet d'apporter de l'information pertinente par rapport aux modèles de l'état de l'art, reposant essentiellement sur des informations de bas niveau (filtres de Gabor) ou sur les positions de points de repère dans le visage. L'intérêt de ces attributs est renforcé par leur capacité à fournir des informations directement interprétables, mais leur principal désavantage est leur difficulté d'extraction. Le calcul automatique de chacun des attributs étant un problème ouvert et faisant actuellement l'objet de nombreuses recherches, il est très probable que les performances de nos modèles augmentent avec l'avancement des techniques d'extraction des attributs.

Nos travaux ont pour objectif d'évaluer différentes photos d'une même personne. Les évaluations ne dépendent pas d'informations sur la beauté de la personne, mais uniquement d'éléments liés à une photo particulière. Les travaux présentés dans ce chapitre nous permettent de proposer une réponse aux questions suivantes :

1. Sur quelle(s) photo(s) cette personne semble-t-elle très sympathique ?
2. Sur quelle(s) photo(s) cette personne semble-t-elle très compétente ?

Pour chacune de ces questions, les algorithmes proposés fournissent des prédictions significativement supérieures au hasard. Les prédictions de sympathie sont même très fortement corrélées aux évaluations de vérité terrain, et nous avons vu que cela pouvait être directement appliqué à la sélection automatique de photographies de visage.

Conclusion et perspectives

Synthèse

L'objectif de ce travail était de proposer différents outils d'analyse de photographies de visage. Nous avons en particulier cherché à évaluer la pertinence d'une photo de visage en fonction d'une application précise : photographies les plus esthétiques, photographies pour lesquelles les personnes représentées véhiculent une impression de sympathie ou de compétence. L'objectif de ces évaluations est de proposer aux utilisateurs des outils permettant de trier, sélectionner ou supprimer automatiquement et rapidement un grand nombre de photographies, ou encore d'améliorer les chances de succès d'une photographie en intégrant nos modèles dans un appareil photo.

Les problèmes liés à la réalisation de tels algorithmes sont nombreux. Tout d'abord, nous cherchons à reproduire un jugement humain subjectif, une même photo peut être considérée comme réussie ou ratée par deux personnes différentes. L'évaluation des performances de ces modèles nécessite la comparaison des évaluations de l'algorithme aux évaluations humaines. Or, ces dernières sont difficiles à obtenir : faire évaluer un grand nombre de photographies par des humains est un processus long et coûteux. Aussi, de très nombreux critères de différente nature sont à prendre en compte pour l'évaluation des impressions dégagées par une photo de visage : expression de la personne, contexte de la photo, qualité intrinsèque de l'image, composition, lumières, couleurs, etc.

Afin de résoudre ces problèmes, nous avons dans un premier temps collecté un grand nombre de photographies de visage, provenant de sources diverses : sites de partage de photographies, bases de photos existantes, photographies privées. Ces photographies ont toutes été évaluées par un nombre significatif d'individus (généralement au-delà de 25 personnes pour chaque image) afin de constituer des bases fiables, dans le sens où les annotations associées aux photographies représentent un niveau de consensus élevé.

Nous avons par la suite étudié les caractéristiques discriminantes permettant de distinguer des photos appropriées à une utilisation donnée de photos non pertinentes. Il ressort par exemple que l'analyse de la netteté dans la région des yeux suffit à prédire efficacement la qualité esthétique de photographies, tandis que la présence d'un sourire et d'un visage jovial est largement corrélée à une impression de sympathie dégagée par un visage.

Ces caractéristiques ont ensuite été utilisées afin de créer des modèles d'estimation de la qualité esthétique, d'impressions de compétence et de sympathie véhiculées par une photo de visage. Si l'évaluation de la compétence peut être améliorée par la prise en compte d'éléments en dehors du visage (vêtements, arrière-plan), nous proposons des modèles d'estimation de la qualité esthétique et de sympathie utilisables dans des situations réelles : les évaluations proposées par l'algorithme sont proches des évaluations humaines. Nous avons en effet mesuré des coefficients de corrélation supérieurs à 0,7 pour ces deux objectifs.

Ces travaux ont fait l'objet de nombreuses comparaisons avec d'autres méthodes d'évaluation de photographies de visage. Dans le cas de la qualité esthétique, nous montrons que les caractéristiques considérées ainsi que les algorithmes d'apprentissage sont particulièrement adaptés aux photos de visage. Nos résultats sont significativement supérieurs aux méthodes existantes avant le début de ce travail. De même, nos modèles d'estimation des impressions de compétence et de sympathie dégagées par une photo de visage sont plus efficaces que les travaux précédents, notamment grâce à l'introduction d'attributs de haut niveau dans les modèles.

Enfin, nous avons proposé un premier exemple de fusion des évaluations de qualité esthétique et d'impression de sympathie. Cette fusion permet de trier les photographies de visage selon deux critères simultanément : les photos de très mauvaise qualité esthétique ou dont les visages semblent très peu sympathiques peuvent être automatiquement supprimées, tandis que les photos de très bonne qualité esthétique présentant des visages très sympathiques peuvent être automatiquement sélectionnées. Cette opération permet par exemple pour une personne donnée de présélectionner des photographies à partager entre amis, en famille, sur un réseau social.

Principales contributions

Nous résumons ici les principales contributions de ce travail.

Dans le chapitre 2, nous avons proposé un cadre de travail général approprié aux différents problèmes étudiés dans ce document. Nous avons adapté un algorithme de sélection de caractéristiques, et fusionné les prédictions de plusieurs algorithmes d'apprentissage. Nous avons montré dans les chapitres 3 et 4 que l'utilisation de ces outils permet d'améliorer la précision et la robustesse des estimations.

Estimation de la qualité esthétique

La principale contribution de notre travail à l'état de l'art de l'estimation de la qualité esthétique de photos de visage est la prise en compte d'informations locales propres aux photos de visage dans nos modèles. Ces informations correspondent au calcul de statistiques globales (netteté, contraste, illumination, couleur) dans les régions correspondant aux yeux et à la bouche. La prise en compte de ces éléments améliore significativement les performances de prédiction des algorithmes.

Nous avons notamment montré que le calcul d'informations dans la région des yeux uniquement permet d'obtenir des performances de prédiction quasiment égales aux résultats obtenus lorsque toutes les régions sont prises en compte. Ce résultat est particulièrement intéressant car il permet d'accélérer significativement le calcul des caractéristiques. Les deux contributions que nous venons d'évoquer ont fait l'objet d'une publication scientifique dans une conférence internationale [Lienhard et al. 2015c].

Estimation des impressions de compétence et de sympathie

Dans ce travail, nous avons montré que l'estimation des impressions de compétence et de sympathie pouvait être significativement améliorée par la prise en compte d'attributs de haut niveau décrivant le visage représenté. Nous avons effectué des expériences montrant que ces attributs étaient plus efficaces que l'utilisation de descripteurs de bas niveau (filtres de Gabor) ou de positions de points de repère dans le visage, actuellement utilisés dans l'état de l'art. Ce résultat est mis en valeur par la rédaction d'une publication scientifique dans une conférence internationale [Lienhard et al. 2015a].

Fusion des modèles de qualité esthétique et d'impression de sympathie

Nous avons enfin cherché à créer un premier modèle tenant compte simultanément de la qualité esthétique et de l'impression de sympathie dégagée par une photo de visage. Les premiers résultats semblent prometteurs et permettent d'envisager des applications de tri ou de sélection automatique de photographies. Nous avons publié ces résultats dans un journal à audience internationale [Lienhard et al. 2015b]. Cet article comprend également les améliorations de performances induites par l'utilisation de l'algorithme Relief et de la fusion des prédictions des différents algorithmes d'apprentissage.

Perspectives

Pour tous les travaux présentés dans ce document, un des facteurs clefs limitant les performances des algorithmes est l'absence de bases d'images contenant à la fois un grand nombre de photos annotées (plusieurs milliers au moins), de sources très variées (photographies amateurs, professionnelles, très réussies ou complètement ratées, etc.) dans un environnement très contrôlé (les évaluations fournies par les sites de partage de photo n'étant pas toujours fiables). Afin de valider définitivement les résultats observés dans ce travail, il sera nécessaire de passer par une phase de création de base de données respectant ces critères. Cette observation vaut pour l'estimation de la qualité esthétique ainsi que pour les estimations d'impressions de compétence et de sympathie.

Estimation de la qualité esthétique

Les caractéristiques que nous proposons sont pertinentes pour l'estimation automatique de la qualité esthétique de photographies de visage. Toutefois, il est possible que certaines informations sur l'esthétisme de l'image soient toujours manquantes. Par exemple, si nous tenons compte d'informations liées au contraste, nous n'avons pas actuellement de mesure permettant d'estimer la distribution de la lumière dans l'image. Or il ressort dans les travaux de [Redi et al. 2015] que cette information est pertinente, les photos de visage dont la qualité esthétique est élevée présentent en effet une luminance plus élevée dans la région du visage. Ce type d'information pourrait améliorer nos scores de prédiction, notamment pour les images de très

bonne qualité esthétique car ce type d'éclairage est généralement utilisé par les photographes professionnels.

Notre méthode peut être appliquée à l'évaluation de photographies contenant des visages, même lorsqu'il existe plusieurs visages dans la photo, ou que le visage détecté n'est pas l'élément central de la photo. Toutefois, les performances de prédiction observées sont alors plus faibles, et il convient de tenir compte d'informations supplémentaires pour améliorer l'estimation de tout type de photographie contenant des visages (portrait d'une personne en entier, photos de groupe, etc.). Ces informations peuvent par exemple correspondre aux relations entre les visages (positions, distances, tailles, orientations). [Li et al. 2010a] montrent par exemple que sur une photo de groupe, il est préférable que les visages soient proches les uns des autres.

De la même manière que les attributs de haut niveau améliorent les performances d'estimation des impressions de compétence et de sympathie, il peut être intéressant d'en considérer certains pour l'estimation de la qualité esthétique. Par exemple, des informations décrivant le contexte dans laquelle la photo est prise (intérieur, extérieur, type d'arrière-plan) peuvent jouer un rôle dans l'estimation de la qualité esthétique. En effet, la position de la ligne d'horizon est un élément important de la qualité esthétique d'une photographie lorsque celle-ci est prise dans un environnement extérieur.

Enfin, nos performances sont tributaires de la précision des algorithmes de détection (ici l'algorithme de Viola-Jones). Utiliser des algorithmes de détection plus performants nous permettrait d'améliorer nos performances d'évaluation, et surtout de traiter les photos dans lesquelles le visage est occulté ou n'est pas présenté de face.

Estimation des impressions de compétence et de sympathie

Comme nous l'avons souligné dans le chapitre 4, de nombreuses informations nécessaires à une prédiction optimale sont toujours manquantes. Afin de proposer un modèle d'estimation plus complet, il est nécessaire d'inclure des informations concernant les vêtements du sujet, les accessoires (bijoux, couvre-chef), le maquillage, et l'arrière-plan.

Une limitation importante des attributs de haut niveau proposés dans ce document est la nécessité d'utiliser des outils d'analyse faciale Betaface, SkyBiometry et SHORE, tous trois propriétaires. Réimplémenter l'extraction des attributs, ou à défaut, utiliser des bibliothèques libres, est une étape nécessaire pour la création d'un outil indépendant et réutilisable.

Liste des publications

Lienhard, Arnaud, Marion Reinhard, Alice Caplier and Patricia Ladret (2014). "Photo Rating of Facial Pictures based on Image Segmentation". In : *9th Int. Conf. on computer Vision Theory and Applications, VISAPP*, p. 329-336.

Lienhard, Arnaud, Patricia Ladret and Alice Caplier (2015). "Low Level Features for Quality Assessment of Facial Images". In : *10th Int. Conf. on computer Vision Theory and Applications, VISAPP*, p. 545-552.

Lienhard, Arnaud, Patricia Ladret and Alice Caplier (2015). "Fully automated facial picture evaluation using high level attributes". In : *11th IEEE International Conference on Automatic Face and Gesture Recognition, FG*, p. 1-6.

Lienhard, Arnaud, Patricia Ladret and Alice Caplier (2015). "How to predict the global instantaneous feeling induced by a facial picture ?". In : *Signal Processing : Image Communication*, Vol. 39 Part C, p. 473-486.

Bibliographie

Aydin, Tunc, Aljoscha Smolic et Markus Gross (2015). « Automated Aesthetic Analysis of Photographic Images ». In : *IEEE Transactions on Visualization and Computer Graphics* 21.1, p. 31–42 (cf. p. 17, 33–35, 44).

Battiato, S, M Moltisanti, F Ravì, AR Bruna et F Naccari (2013). « Aesthetic scoring of digital portraits for consumer applications ». In : p. 866008–866008 (cf. p. 19, 22, 44).

Bay, Herbert, Tinne Tuytelaars et Luc Van Gool (2006). « Surf : Speeded up robust features ». In : *Computer vision–ECCV 2006*. Springer, p. 404–417 (cf. p. 14).

Betaface. URL : `http://www.Betaface.com/` (cf. p. 29).

Bhattacharya, Subhabrata, Rahul Sukthankar et Mubarak Shah (2010). « A framework for photo-quality assessment and enhancement based on visual aesthetics ». In : *Proceedings of the international conference on Multimedia - MM '10*, p. 271 (cf. p. 35).

Breiman, Leo (2001). « Random forests ». In : *Machine learning* 45, p. 5–32 (cf. p. 40, 58, 64).

Breiman, Leo, Jerome Friedman, Charles J Stone et Richard A Olshen (1984). *Classification and regression trees*. CRC press (cf. p. 64).

Cerosaletti, CD et AC Loui (2009). « Measuring the perceived aesthetic quality of photographic images ». In : *Quality of Multimedia Experience*, p. 47–52 (cf. p. 24, 26).

Chang, CC et CJ Lin (2011). « LIBSVM : a library for support vector machines ». In : *ACM Transactions on Intelligent Systems and Technologies*, p. 1–39 (cf. p. 39, 59).

Cohen-Or, Daniel, Olga Sorkine, Ran Gal, Tommer Leyvand et Ying-Qing Xu (2006). « Color harmonization ». In : *ACM SIGGRAPH*, p. 624 (cf. p. 37).

Crete, Frederique, Thierry Dolmiere, Patricia Ladret et Marina Nicolas (2007). « The blur effect : perception and estimation with a new no-reference perceptual blur metric ». In : *Electronic Imaging 2007*. International Society for Optics et Photonics, p. 64920I–64920I (cf. p. 32, 89).

Dalal, Navneet et Bill Triggs (2005). « Histograms of oriented gradients for human detection ». In : *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. T. 1. IEEE, p. 886–893 (cf. p. 28).

Dantcheva, Antitza, Carmelo Velardo, Angela D'angelo et Jean-Luc Dugelay (2011). « Bag of soft biometrics for person identification ». In : *Multimedia Tools and Applications* 51.2, p. 739–777 (cf. p. 130).

Datta, R, J Li et JZ Wang (2008). « Algorithmic inferencing of aesthetics and emotion in natural images : An exposition ». In : *Image Processing, ICIP*, p. 105–108 (cf. p. 18).

Datta, Ritendra et JZ Wang (2010). « ACQUINE : aesthetic quality inference engine-real-time automatic rating of photo aesthetics ». In : *Proceedings of the international conference on Multimedia*, p. 1–4 (cf. p. 18, 19).

Datta, Ritendra, D Joshi, J Li et JZ Wang (2006). « Studying aesthetics in photographic images using a computational approach ». In : *Computer Vision–ECCV 2006*, p. 288–301 (cf. p. 14, 16, 18, 23, 24, 33–37, 39, 41, 97).

Datta, Ritendra, Jia Li et James Z. Wang (2007). « Learning the consensus on visual quality for next-generation image management ». In : *Proceedings of the 15th international conference on Multimedia*. 2. New York, New York, USA : ACM Press, p. 533 (cf. p. 17, 38).

Desnoyer, Mark et David Wettergreen (2010). « Aesthetic Image Classification for Autonomous Agents ». In : *20th International Conference on Pattern Recognition*, p. 3452–3455 (cf. p. 24, 25, 32–35, 37, 38).

Dhar, Sagnik, V Ordonez et TL Berg (2011). « High level describable attributes for predicting aesthetics and interestingness ». In : *Computer Vision and Pattern Recognition*, p. 1657–1664 (cf. p. 15, 24, 25, 34–37, 39).

DPChallenge. URL : `http://www.dpchallenge.com/` (cf. p. 12, 20, 22, 24, 25, 79, 83, 84, 87).

Drucker, Harris, CJC Burges, Linda Kaufman, Alex Smola et Vladimir Vapnik (1997). « Support vector regression machines ». In : *Neural Information Processing Systems*, p. 155–161 (cf. p. 58).

Ernst, Andreas, Tobias Ruf et Christian Kueblbeck (2009). « A modular framework to detect and analyze faces for audience measurement systems ». In : *2nd Workshop on Pervasive Advertising at Informatik*, p. 75–87 (cf. p. 20).

FaceGen. URL : `http://www.facegen.com/` (cf. p. 30, 130, 131).

FacePlusPlus. URL : `http://www.faceplusplus.com/` (cf. p. 22).

Faria, J, S Bagley, Stefan Rüger et Toby Breckon (2013). « Challenges of finding aesthetically pleasing images ». In : *Image Analysis for Multimedia Interactive Services (WIAMIS)*. T. 2, p. 4–7 (cf. p. 16, 38, 39).

Fedorovskaya, EA (2002). « Perceived overall contrast and quality of the tone scale rendering for natural images ». In : *Electronic Imaging* 4662, p. 119–128 (cf. p. 33, 35, 38).

Fiche, Cécile, Patricia Ladret et Ngoc-Son Vu (2010). « Blurred face recognition algorithm guided by a no-reference blur metric ». In : *IS&T/SPIE Electronic Imaging*. International Society for Optics et Photonics, 75380U–75380U (cf. p. 11).

Flickr. URL : `http://www.flickr.com/` (cf. p. 23, 24, 26, 85, 87, 102, 117).

Freund, Yoav, Robert Schapire et N Abe (1999). « A short introduction to boosting ». In : *Journal-Japanese Society For Artificial Intelligence* 14.771-780, p. 1612 (cf. p. 98).

Friedman, Jerome H (2001). « Greedy function approximation : a gradient boosting machine ». In : *Annals of statistics*, p. 1189–1232 (cf. p. 58, 66).

Hasler, David et SE Suesstrunk (2003). « Measuring colorfulness in natural images ». In : *Electronic Imaging. International Society for Optics and Photonics*. P. 87–95 (cf. p. 34, 95).

He, Kaiming, Jian Sun et Xiaoou Tang (2011). « Single image haze removal using dark channel prior ». In : *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33.12, p. 2341–2353 (cf. p. 95).

Huang, GB, M Mattar, Tamara Berg et E Learned-Miller (2007). *Labeled faces in the wild : A database for studying face recognition in unconstrained environments*. Rapp. tech., p. 1–11 (cf. p. 81, 101).

Itti, Laurent, Christof Koch et Ernst Niebur (1998). « A model of saliency-based visual attention for rapid scene analysis ». In : *IEEE Transactions on pattern analysis and machine intelligence* 20.11, p. 1254–1259 (cf. p. 15, 35).

Jiang, Wei, Alexander C. Loui et Cathleen Daniels Cerosaletti (2010). « Automatic aesthetic value assessment in photographic images ». In : *IEEE International Conference on Multimedia and Expo*, p. 920–925 (cf. p. 19, 24, 26, 32, 33, 35–37, 39).

Joshi, D et al. (2011). « Aesthetics and emotions in images ». In : *Signal Processing Magazine* SEPTEMBER 2011, p. 94–115 (cf. p. 18).

Ke, Yan, Xiaoou Tang et Feng Jing (2006). « The design of high-level features for photo quality assessment ». In : *Computer Vision and Pattern Recognition*. T. 1, p. 419–426 (cf. p. 11–16, 19, 24, 25, 32–34, 37, 38).

Khan, SS et Daniel Vogel (2012). « Evaluating visual aesthetics in photographic portraiture ». In : *Computational Aesthetics in Graphics, Visualization and Imaging*, p. 1–8 (cf. p. 20–22, 24, 32, 33, 35–39, 80, 86, 118).

Kim, Jonghee et Changick Kim (2014). « Aesthetic Quality Classification via Subject Region Extraction ». In : *International Conference on Image Processing (ICIP)*, p. 536–540 (cf. p. 16, 24, 25, 32–36, 39, 97, 115).

Kononenko, Igor (1994). « Estimating attributes : analysis and extensions of RELIEF ». In : *Machine Learning : ECML-94*. Springer, p. 171–182 (cf. p. 54).

Kumar, Neeraj, Alexander C Berg, Peter N Belhumeur et Shree K Nayar (2009). « Attribute and simile classifiers for face verification ». In : *EEE 12th International Conference on Computer Vision*, p. 365–372 (cf. p. 130).

Lajevardi, Seyed Mehdi et Margaret Lech (2008). « Averaged gabor filter features for facial expression recognition ». In : *Digital Image Computing : Techniques and Applications (DICTA), 2008*. IEEE, p. 71–76 (cf. p. 28, 32, 135, 136, 142).

LeCun, Yann, Leon Bottou, Yoshua Bengio et Patrick Haffner (1998). « Gradient-based learning applied to document recognition ». In : *Proceedings of the IEEE* 86.11, p. 2278–2324 (cf. p. 44, 58).

Li, Congcong et Tsuhan Chen (2009). « Aesthetic visual quality assessment of paintings ». In : *IEEE Selected Topics in Signal Processing* 3.2, p. 236–252 (cf. p. 18, 33).

Li, Congcong, Andrew Gallagher, Alexander C. Loui et Tsuhan Chen (2010a). « Aesthetic quality assessment of consumer photos with faces ». In : *International Conference on Image Processing (ICIP)*, p. 3–6 (cf. p. 5, 19, 20, 24, 35, 37, 39, 41, 80, 85, 87, 88, 117–121, 125, 158).

Li, Congcong, AC Loui et T Chen (2010b). « Towards aesthetics : a photo quality assessment and photo selection system ». In : *Proceedings of the international conference on Multimedia*, p. 10–13 (cf. p. 18, 19).

Li, Haoxiang, Zhe Lin, Xiaohui Shen, Jonathan Brandt et Gang Hua (2015). « A Convolutional Neural Network Cascade for Face Detection ». In : *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 5325–5334 (cf. p. 97).

Liao, Pin et al. (2012). « Facial Image Quality Assessment Based on Support Vector Machines ». In : *International Conference on Biomedical Engineering and Biotechnology*, p. 810–813 (cf. p. 11).

Lienhard, Arnaud, Patricia Ladret et Alice Caplier (2015a). « Fully automated facial picture evaluation using high level attributes ». In : *11th IEEE Int. Conf. Automatic Face and Gesture Recognition*, p. 1–6 (cf. p. 157).

Lienhard, Arnaud, Patricia Ladret et Alice Caplier (2015b). « How to predict the global instantaneous feeling induced by a facial picture ? » In : *Signal Processing : Image Communication* 39-C, p. 473–486 (cf. p. 157).

— (2015c). « Low Level Features for Quality Assessment of Facial Images ». In : *10th Int. Conf. on computer Vision Theory and Applications, VISAPP*, p. 545–552 (cf. p. 156).

Lienhart, R. et J. Maydt (2002). « An extended set of Haar-like features for rapid object detection ». In : *International Conference on Image Processing (ICIP)*. T. 1. Ieee, p. I–900–I–903 (cf. p. 98, 99).

Liu, Lixiong, Bao Liu, Hua Huang et Alan Conrad Bovik (2014). « No-reference image quality assessment based on spatial and spectral entropies ». In : *Signal Processing : Image Communication* 29.8, p. 856–863 (cf. p. 11).

Loui, A, MD Wood, Anthony Scalise et John Birkelund (2008). « Multidimensional image value assessment and rating for automated albuming and retrieval ». In : *Image Processing, ICIP*, p. 97–100 (cf. p. 35, 39).

Lowe, David G (1999). « Object recognition from local scale-invariant features ». In : *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. T. 2. Ieee, p. 1150–1157 (cf. p. 14).

Luo, Wei, Xiaoou Tang et Xiaogang Wang (2011). « Content-Based Photo Quality Assessment ». In : *IEEE Transactions on Multimedia* 15.8, p. 1930–1943 (cf. p. 95).

Luo, Yiwen et Xiaoou Tang (2008). « Photo and video quality evaluation : Focusing on the subject ». In : *Computer Vision–ECCV 2008*, p. 386–399 (cf. p. 12, 13, 16, 18, 19, 24, 33–35, 37–39, 97).

Ma, Yu-Fei, Lie Lu, Hong-Jiang Zhang et Mingjing Li (2002). « A user attention model for video summarization ». In : p. 533–542 (cf. p. 12).

Males, M, A Hedi et M Grgic (2013). « Aesthetic quality assessment of headshots ». In : *55th International Symposium ELMAR*. September, p. 25–27 (cf. p. 21, 24, 32–35, 39, 41).

Marat, Sophie, Anis Rahman, Denis Pellerin, Nathalie Guyader et Dominique Houzet (2013). « Improving visual saliency by adding face feature map and center bias ». In : *Cognitive Computation* 5.1, p. 63–75 (cf. p. 13).

Marchesotti, Luca et Florent Perronnin (2011). « Assessing the aesthetic quality of photographs using generic image descriptors ». In : *IEEE International Conference on Computer Vision (ICCV)*, p. 1784–1791 (cf. p. 14, 23, 24, 32, 35, 39, 44, 78).

Mazza, Filippo, Matthieu Perreira Da Silva, Patrick Le Callet et IEJ Heynderickx (2015). « What do you think of my picture ? Investigating factors of influence in profile images context perception ». In : *IS&T/SPIE Electronic Imaging*. International Society for Optics et Photonics, p. 93940D–93940D (cf. p. 2, 130, 134).

Muja, Marius et David G Lowe (2009). « Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. » In : *VISAPP (1)* 2 (cf. p. 52).

Murray, Naila, Luca Marchesotti et Florent Perronnin (2012). « AVA : A large-scale database for aesthetic visual analysis ». In : *Computer Vision and Pattern Recognition*, p. 2408–2415 (cf. p. 25, 83).

Ng, Wai-Seng et al. (2009). « Automatic Photo Ranking Based on Esthetics Rules of Photography ». In : *Proceedings of Computer Graphics Workshop* (cf. p. 15, 16, 18, 24, 25, 33, 35–37, 39).

Nishiyama, Masashi, Takahiro Okabe, Yoichi Sato et Imari Sato (2009). « Sensation-based photo cropping ». In : *Proceedings of the seventeen ACM international conference on Multimedia*, p. 669 (cf. p. 18, 34).

Oliva, Aude et A Torralba (2001). « Modeling the shape of the scene : A holistic representation of the spatial envelope ». In : *International journal of computer vision* 42.3, p. 145–175 (cf. p. 32).

Oosterhof, Nikolaas N et Alexander Todorov (2008). « The functional basis of face evaluation. » In : *Proceedings of the National Academy of Sciences of the United States of America* 105.32, p. 11087–92 (cf. p. 27, 30).

PhotoNet. URL : `http://www.photo.net/` (cf. p. 8, 14, 23, 24, 79–81, 87).

Pogačnik, D, Robert Ravnik, Narvika Bovcon et Franc Solina (2012). « Evaluating photo aesthetics using machine learning ». In : *Data Mining and Data Warehouses*, p. 4–7 (cf. p. 20, 21, 24, 25, 34, 35, 37, 39, 50, 117).

Ravì, Fabrizio et Sebastiano Battiato (2012). « A novel computational tool for aesthetic scoring of digital photography ». In : *Conference on Colour in Graphics, Imaging and Vision* 2012.1, p. 349–354 (cf. p. 35, 37).

Redi, Miriam, Nikhil Rasiwasia, Gaurav Aggarwal et Alejandro Jaimes (2015). « The Beauty of Capturing Faces : Rating the Quality of Digital Portraits ». In : *Automatic Face and Gesture Recognition*, p. 43. arXiv : `1501.0730` (cf. p. 21, 22, 24, 25, 32–37, 39, 41, 80, 86–88, 117, 119–121, 125, 157).

Riaz, Sidra, KH Lee et SW Lee (2012). « Aesthetic score assessment based on generic features in digital photography ». In : *5th AUN/SEED-Net Regional Conference on Information and Communicaation Technology*, p. 76–79 (cf. p. 32–35, 39).

Riedmiller, M. et H. Braun (1993). « A direct adaptive method for faster backpropagation learning : the RPROP algorithm ». In : *IEEE International Conference on Neural Networks*. Ieee, p. 586–591 (cf. p. 58, 61).

Robnik-Šikonja, Marko et Igor Kononenko (1997). « An adaptation of Relief for attribute estimation in regression ». In : *Machine Learning : Proceedings of the Fourteenth International Conference (ICML'97)*, p. 296–304 (cf. p. 57).

Robnik-Šikonja, M et I Kononenko (2003). « Theoretical and empirical analysis of ReliefF and RReliefF ». In : *Machine learning* 53, p. 23–69 (cf. p. 21, 50, 52, 54, 55).

Rojas, Mario, David Masip, Alexander Todorov, Jordi Vitrià et al. (2010). « Automatic point-based facial trait judgments evaluation ». In : *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, p. 2715–2720 (cf. p. 130, 131, 142, 147, 148).

Rojas, Mario, David Masip, Alexander Todorov, Jordi Vitria et al. (2011). « Automatic prediction of facial trait judgments : Appearance vs. structural models ». In : *PloS one* 6.8, p. 1–12 (cf. p. 28, 32, 36, 130, 131, 142).

Rumelhart, DE, GE Hinton et RJ Williams (1988). « Learning representations by back-propagating errors ». In : *Cognitive modeling* 323 (cf. p. 61).

SHORE. URL : `http://www.iis.fraunhofer.de/en/ff/bsy/dl/shore` (cf. p. 29).

Siagian, Christian et Laurent Itti (2007). « Rapid biologically-inspired scene classification using features shared with visual attention. » In : *IEEE transactions on pattern analysis and machine intelligence* 29.2, p. 300–12 (cf. p. 32, 35).

SkyBiometry. URL : `http://www.SkyBiometry.com/` (cf. p. 29).

Sun, Xiaoshuai, Hongxun Yao, Rongrong Ji et Shaohui Liu (2009). « Photo assessment based on computational visual attention model ». In : *Proceedings of the 17th ACM international conference on mumtimedia*. 92, p. 541–544 (cf. p. 35).

Tang, Xiaoou, Wei Luo et Xiaogang Wang (2013). « Content-based photo quality assessment ». In : *Transactions on Multimedia* 15.8, p. 1930–1943 (cf. p. 16–18, 24, 25, 34–37, 39, 84, 87, 115).

Todorov, Alexander et NN Oosterhof (2011). « Modeling Social Perception of Faces ». In : *Signal Processing Magazine, IEEE* March, p. 117–122 (cf. p. 27, 28, 30, 36, 131).

Todorov, Alexander, Chris P Said, Andrew D Engell et Nikolaas N Oosterhof (2008). « Understanding evaluation of faces on social dimensions ». In : *Trends in cognitive sciences* 12.12, p. 455–460 (cf. p. 27, 132, 133).

Todorov, Alexander, Ron Dotsch, Jenny M Porter, Nikolaas N Oosterhof et Virginia B Falvello (2013). « Validation of data-driven computational models of social perception of faces. » In : *Emotion* 13.4, p. 724–38 (cf. p. 28, 30, 132).

Tong, Hanghang, Mingjing Li et Hongjiang Zhang (2004). « Blur detection for digital images using wavelet transform ». In : *IEEE International Conference on Multimedia and Expo (ICME)*. Ieee, p. 17–20 (cf. p. 11–13, 25, 33–35, 38, 39).

Vapnik, Vladimir Naumovich et Vlamimir Vapnik (1998). *Statistical learning theory*. T. 1. Wiley New York (cf. p. 57).

Viola, P. et M. Jones (2001). « Rapid object detection using a boosted cascade of simple features ». In : *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. T. 1. IEEE Comput. Soc, p. I–511–I–518 (cf. p. 19, 82, 97, 98, 103).

Willis, Janine et Alexander Todorov (2006). « Making Up Your Mind After a 100-Ms Exposure to a Face ». In : *Psychological science* 17.7, p. 592–598 (cf. p. 1, 26).

Wong, Lai-kuan et Kok-lim Low (2009). « Saliency-enhanced image aesthetics class prediction ». In : *16th IEEE International Conference on Image Processing*. Ieee, p. 997–1000 (cf. p. 15, 16, 19, 23, 24, 33, 35, 39).

Xue, SF, Henry Tang et Dan Tretter (2013). « Feature design for aesthetic inference on photos with faces ». In : *Image Processing (ICIP)*, p. 2689–2693 (cf. p. 5, 19, 24, 32–35, 37, 39, 85, 120, 121, 125).

Résumé — Avec le développement des appareils photos numériques et des sites de partage de photos, nous passons une part croissante de notre temps à observer, sélectionner et partager des images, parmi lesquelles figurent un grand nombre de photos de visage. Dans cette thèse, nous nous proposons de créer un premier système entièrement automatique renvoyant une estimation de la pertinence d'une photo de visage pour son utilisation dans la création d'un album de photos, la sélection de photos pour un réseau social ou professionnel, etc. Pour cela, nous créons plusieurs modèles d'estimation de la pertinence d'une photo de visage en fonction de son utilisation. Dans un premier temps, nous adaptons les modèles d'estimation de la qualité esthétique d'une photo au cas particulier des photos de visage. Nous montrons que le fait de calculer 15 caractéristiques décrivant différents aspects de l'image (texture, illumination, couleurs) dans des régions spécifiques de l'image (le visage, les yeux, la bouche) améliore significativement la précision des estimations par rapport aux modèles de l'état de l'art. La précision de ce modèle est renforcée par la sélection de caractéristiques adaptées à notre problème, ainsi que par la fusion des prédictions de 4 algorithmes d'apprentissage. Dans un second temps, nous proposons d'enrichir l'évaluation automatique d'une photo de visage en définissant des modèles d'estimation associés à des critères tels que le degré de sympathie ou de compétence dégagé par une photo de visage. Ces modèles reposent sur l'utilisation d'attributs de haut niveau (présence de sourire, ouverture des yeux, expressions faciales), qui se montrent plus efficaces que les caractéristiques de bas niveau utilisées dans l'état de l'art (filtres de Gabor, position des points de repère du visage). Enfin, nous fusionnons ces modèles afin de sélectionner automatiquement des photos de bonne qualité esthétique et appropriées à une utilisation donnée : photos inspirant de la sympathie à partager en famille, photos dégageant une impression de compétence sur un réseau professionnel.

Mots clés : Portrait, Qualité esthétique, Compétence, Sympathie, Évaluation.

Abstract — Picture selection is a time-consuming task for humans and a real challenge for machines, which have to retrieve complex and subjective information from image pixels. An automated system that infers human feelings from digital portraits would be of great help for profile picture selection, photo album creation or photo editing. In this work, several models of facial pictures evaluation are defined. The first one predicts the overall aesthetic quality of a facial image by computing 15 features that encode low-level statistics in different image regions (face, eyes and mouth). Relevant features are automatically selected by a feature ranking technique, and the outputs of 4 learning algorithms are fused in order to make a robust and accurate prediction of the image quality. Results are compared with recent works and the proposed algorithm obtains the best performance. The same pipeline is then considered to evaluate the likability and competence induced by a facial picture, with the difference that the estimation is based on high-level attributes such as gender, age and smile. Performance of these attributes is compared with previous techniques that mostly rely on facial keypoints positions, and it is shown that it is possible to obtain predictions that are close to human perception. Finally, a combination of both models that selects a likable facial image of good aesthetic quality for a given person is described.

Keywords : Portrait, Aesthetic quality, Competence, Likability, Assessment.